AVA-ENCODER: TOWARDS AGENT-NATIVE VIDEO REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic video creation systems remain far from cinematic-grade quality: their underlying models struggle to plan how story, characters, visuals, and sound fit together, and high-quality creation trajectories to learn such planning abilities are scarce. Finished human films embody this knowledge, but in a tightly coupled audiovisual form that agents cannot directly read or manipulate. We propose the Agentic Video Auto-Encoder (AVA-Encoder), which works backward from finished films, turning them into agent-native representations with reconstruction itself as supervision. It encodes a film into a Film Knowledge Graph (Film KG) of structured-text nodes, typed dependency edges, and linked multimodal assets, then regenerates the film from this graph with a fixed decoder. Reconstruction residuals are distilled into textual gradients that drive dual-loop self-evolution: an outer loop learns a shared encoding policy across videos, and an inner loop refines each video's Film KG at test time. AVA-Encoder improves reconstruction fidelity by 20.7 percentage points over the strongest baseline; its learned policy outperforms a finely human-tuned policy with over 70% fewer system-prompt tokens; and Film KGs support coordinated editing and improve all tested downstream creation systems. We release the framework, a reconstruction benchmark, and a Film KG dataset.

1 INTRODUCTION

Advances in foundation models (Brooks et al., 2024; Google DeepMind, 2024; Kuaishou Technology, 2024; Polyak et al., 2024) and agentic video creation systems (Xu et al., 2025; Wu et al., 2025; Li et al., 2024; Lin et al., 2024; Lai et al., 2026) have enabled agents to write stories, design keyframes, and generate videos. Yet reliable cinematic-grade creation remains difficult, mainly because their underlying models lack the planning ability to coordinate scripts, characters, shots, and audiovisual elements. Developing this ability is difficult because high-quality trajectories of complete agentic video creation processes are scarce. Meanwhile, cinematic-grade human-made films offer a rich alternative source of filmmaking knowledge. However, unlike human filmmakers who learn by directly studying finished films, agents cannot directly use them as explicit creation records.

This limitation reflects a mismatch between video space and agent space: films tightly couple narrative, visual, temporal, and audio information without exposing their underlying relationships (Huang et al., 2020), whereas agents use structured text, code, plans, and graphs for retrieval, reasoning, planning, and editing (Wang et al., 2025). Bridging this gap requires agent-native representations that make film content and creation dependencies explicit and operable.

An ideal agent-native video representation should satisfy three requirements: understandable to agents, stably-editable for agents, and faithful enough to preserve the cinematic information for subsequent generation. Existing approaches do not fully meet these requirements: visual embeddings resist direct interpretation and manipulation; textual descriptions often omit cross-modal dependencies; and fact-oriented structured representation prioritize task-relevant facts over the cinematic detail needed for generation (Sec. 2.2).

Therefore, we propose the Film Knowledge Graph (Film KG) representation, which organizes cinematic content into structured textual nodes linked to generated image, audio, and video assets. Textual nodes make content readable, learnable, and individually editable, while explicit edges preserve dependencies across modalities and levels for reconstruction and coordinated editing.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To learn faithful Film KGs, we introduce AVA-Encoder, an agentic video auto-encoding framework driven by self-evolution. It encodes a film into a Film KG and reconstructs it through a fixed generative decoder. Reconstruction residuals provide self-supervision for representational faithfulness (He et al., 2022; Mildenhall et al., 2020), guiding self-evolution of a shared encoding policy and an input-specific Film KG at separate stages (Figure 2), without ground-truth graphs or expert creation trajectories.

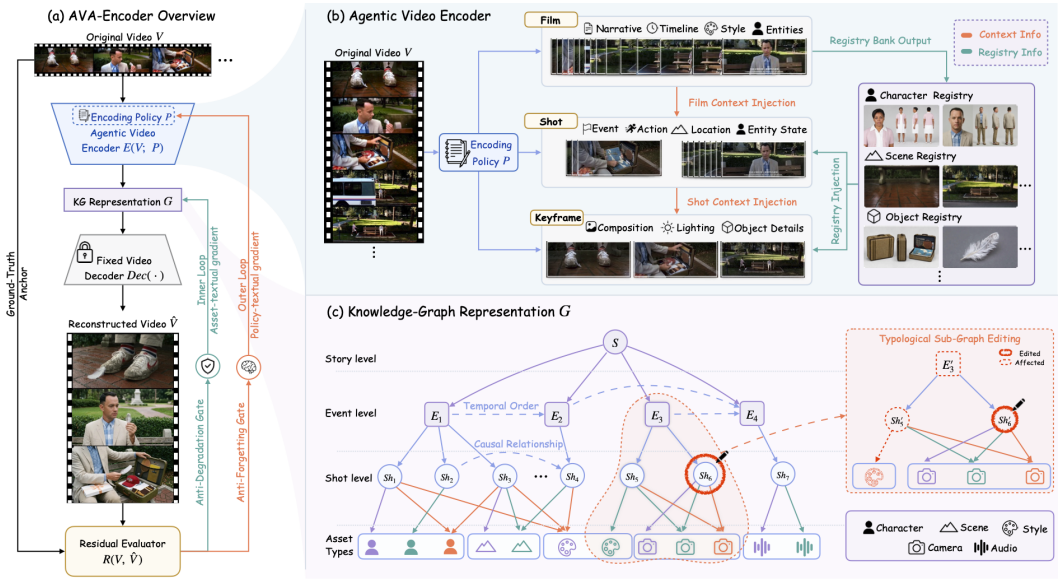


Figure 1: Overview of AVA-Encoder. (a) Reconstruction-driven, gated dual-loop self-evolution. (b) Hierarchical agentic encoding with inter-level context injection. (c) Film KG nodes, edges, and linked multimodal assets for constrained generation and coordinated editing.

Learning such representations poses three challenges, addressed by AVA-Encoder's complementary designs. First, translating dense multimodal content into structured text risks losing cinematic detail. The Agentic Video Encoder (Sec. 4.1) combines film-, shot-, and keyframe-level understanding with hierarchical context injection to preserve local details and global dependencies. Second, reconstruction and consistent editing require preserving dependencies across narrative hierarchy, temporal order, and audiovisual content. Film KG (Sec. 4.2) makes these explicit through typed edges linking textual nodes and assets. Third, stable self-evolution requires controllable updates, precise optimization signals, and robust evaluation. AVA-Encoder uses gated dual-loop textual-gradient optimization (Sec. 4.3) to pseudo-train a generalizable encoding policy and refine the input-specific Film KG at test time. Acceptance gates limit regressions, fact-level reconstruction feedback guides targeted corrections, and separate optimization and final evaluation signals mitigate reward exploitation (Sec. 4.3.1).

Our contributions are threefold:

1. **Agentic Video Auto-Encoder.** We introduce the Agentic Video Auto-Encoder paradigm, formulating agent-native video representation learning as a self-evolving agentic auto-encoding problem. It achieves 49.0% Overall reconstruction fidelity, exceeding the strongest external baseline by 20.7 percentage points. Its pseudo-trained encoding policy surpasses a human-tuned policy by 1.4 points in Overall fidelity with over 70% fewer system-prompt tokens.
2. **Agentic Video Representation Benchmark.** We introduce an agentic video reconstruction benchmark spanning four evaluation directions and eight cinematic dimensions, with 97.3% agreement between automatic and human pairwise rankings under blinded evaluation.
3. **Film KG Dataset and Editing Framework.** We release high-quality structured-text Film KGs as reverse-engineered creation records, easing the trajectory scarcity noted above, and a framework for coordinated graph-based editing. Film KG representations improve Overall generation quality across all tested open-source downstream agentic video creation systems.

2 Related Work

2.1 Agentic Video Creation Systems

LLM- and VLM-based systems use planning and tool orchestration for video generation (Lin et al., 2024; Zhuang et al., 2024; Xie et al., 2024; Yuan et al., 2024; Lai et al., 2026; Li et al., 2024; Long et al., 2024; Wu et al., 2025; Xu et al., 2025), editing, and remixing (Song et al., 2026; comfyanonymous and ComfyUI contributors, 2023). However, their foundation models struggle to plan complex multimodal cinematic tasks, limiting creation-trajectory quality; the scarcity of high-quality trajectories in this field, in turn, hinders further training to improve planning abilities.

AVA-Encoder instead works backward from human films, converting them into readable, learnable, and editable Film KGs with explicit dependencies, and self-improving the system through reconstruction. Decoding produces aligned intermediate records and rendered outputs, serving as reverse-engineered cinematic-grade creation trajectories for future training, providing a route to learning from abundant high-quality films beyond trajectories current agentic systems can produce.

2.2 Agent-Native Video Representations

Existing video representations include low-level visual forms, such as pixels, latent tokens, and embeddings (Tong et al., 2022; Yu et al., 2023; Zhang et al., 2023); textual descriptions, such as captions and screenplays (Huang et al., 2020; Chung & Yu, 2023; Mahon & Lapata, 2024); and structured representations, such as hierarchical memories and graphs (Song et al., 2024; He et al., 2024; Ataallah et al., 2024; Chu et al., 2025; Rege et al., 2026).

Jointly supporting agent readability, reliable editing, and faithful generation remains challenging. Low-level forms retain visual information but are difficult for agents to read, query, or modify directly; textual descriptions are accessible but often leave entity relationships and cross-modal dependencies implicit. Structured representations support factual retrieval and reasoning, yet need not preserve the cinematic details and generation dependencies required for reconstruction. For example, recording that two characters meet in shot eight does not specify their appearance, spatial arrangement, or audiovisual context required for generation.

Success in retrieval and QA (Zhang et al., 2024a;b; Qian et al., 2024) therefore does not establish generative fidelity. Reconstruction tests whether retained information supports faithful regeneration; AVA-Encoder uses this source-referenced feedback to assess and improve representation fidelity.

3 Task Definition

3.1 Problem Formulation

We formulate agent-native video representation learning as agentic auto-encoding, where an agent autonomously learns to encode a video into a structured agentic representation from which a fixed decoder faithfully reconstructs the video. Given a cinematic video input $V \in \mathcal{V}$ and an encoding policy $P \in \mathcal{P}$, the agentic encoder $E(\cdot)$ and fixed decoder $\mathrm{Dec}(\cdot)$ define:

$$G = E(V; P), \qquad \hat{V} = \mathrm{Dec}(G). \tag{1}$$

Here, the output of the encoder $G \in \mathcal{G}$ is a Film KG. It represents the input film through structured-text nodes describing cinematic content, typed edges encoding dependencies among these elements, and linked multimodal assets synthesized from the textual nodes (Sec. 4.2). All source-video information used for reconstruction is mediated by the textual graph. Unlike dense neural embeddings, this explicit representation G is designed to meet the three agent-native requirements in Sec. 1. $\mathcal{V}$, $\mathcal{P}$, and $\mathcal{G}$ denote the video, encoding policy, and Film KG spaces, respectively.

3.2 Optimization Objective

Under the fixed decoder, reconstruction fidelity measures how well G preserves the source video. Let $R(V, \hat{V})$ denote reconstruction discrepancies and $\mathcal{L}(\cdot)$ map them to a scalar loss:

$$\ell(V, G) := \mathcal{L}(R(V, \mathrm{Dec}(G))) \,. \tag{2}$$

From $G = E(V; P)$, this loss evaluates both an individual representation and the encoding policy that produces it. We therefore formulate optimization at both the policy and instance levels:

$$\begin{cases} \min\limits_{P \in \mathcal{P}} \ \mathbb{E}_{V' \sim \mathrm{Unif}(\mathcal{D})} \left[\ell(V', E(V'; P)) \right], & \text{policy level,} \\ \min\limits_{G \in \mathcal{G}} \ \ell(V, G), & \text{instance level.} \end{cases} \tag{3}$$

where $\mathcal{D} = \{V_i\}_{i=1}^{L}$ is the policy-learning collection of L videos. The policy-level objective averages reconstruction loss over $V' \sim \mathrm{Unif}(\mathcal{D})$, while the instance-level objective minimizes it for a given V. AVA-Encoder realizes these objectives through shared-policy pseudo-training and optional input-specific Film KG refinement within the dual-loop textual-gradient framework (Sec. 4.3). Here, *pseudo-training* fits P to $\mathcal{D}$ from loss feedback and evaluates it on held-out videos as in training, but updates only the natural-language policy, with model weights frozen.

4 METHOD: AVA-ENCODER

AVA-Encoder addresses the three challenges in Sec. 1 by building an agentic auto-encoding framework consisting of a hierarchical agentic video encoder (Sec. 4.1), a Film KG representation (Sec. 4.2), and reconstruction-driven dual-loop textual-gradient evolution (Sec. 4.3; Figure 1(a)).

4.1 AGENTIC VIDEO ENCODER

Mapping a high-dimensional film into structured text risks losing fine-grained cinematic information (Sec. 1). The Agentic Video Encoder addresses this through *three-level specialized understanding* across film, shot, and keyframe granularities, together with *inter-level context injection* that preserves dependencies across these levels (Figure 1(b)). After preprocessing, $E(V; P)$ first extracts hierarchical structured descriptions and relations, then assembles them into the Film KG G.

We first partition a video V into S ordered shots $(s_i)_{i=1}^{S}$ and select J_i keyframes $(f_{i,j}^{\mathrm{kf}})_{j=1}^{J_i}$ from each shot (Appendix Sec. E.1).

Step 1: Hierarchical Agentic Understanding. To capture both local cinematic details and global dependencies, a hierarchical encoding policy $P = (P_{\mathrm{film}}, P_{\mathrm{shot}}, P_{\mathrm{kf}})$ guides a video-understanding agent to extract textual information $\mathcal{C}_{\mathrm{film}}, \mathcal{C}_{\mathrm{shot},i}, \mathcal{C}_{\mathrm{kf},i,j}$ through $\mathrm{Understand}$ at film, shot, and keyframe levels while conditioning on broader cross-hierarchy context:

$$\begin{aligned} (\mathcal{C}_{\mathrm{film}}, \mathcal{B}_{\mathrm{reg}}) &= \mathrm{Understand}(V; P_{\mathrm{film}}), \\ \mathcal{C}_{\mathrm{shot},i} &= \mathrm{Understand}(s_i, \mathcal{C}_{\mathrm{film}}, \mathcal{B}_{\mathrm{reg}}; P_{\mathrm{shot}}), \\ \mathcal{C}_{\mathrm{kf},i,j} &= \mathrm{Understand}(f_{i,j}^{\mathrm{kf}}, \mathcal{C}_{\mathrm{shot},i}, \mathcal{B}_{\mathrm{reg}}; P_{\mathrm{kf}}). \end{aligned} \tag{4}$$

The encoding policy P is optimized through outer-loop pseudo-training (Sec. 4.3.2).

Three-Level Specialized Understanding. Film-level understanding captures global narrative and cross-shot relations while constructing a shared registry $\mathcal{B}_{\mathrm{reg}}$ of canonical entity names and film-wide identifiers. Shot-level understanding captures temporal, audiovisual, and camera dynamics, whereas keyframe-level understanding extracts fine-grained visual composition and appearance.

Inter-Level Information Injection. Broader context is injected into finer levels to preserve cross-level dependency and consistency: $\mathcal{C}_{\mathrm{film}}$ conditions shot understanding, $\mathcal{C}_{\mathrm{shot},i}$ conditions keyframe understanding, and the shared registry $\mathcal{B}_{\mathrm{reg}}$ maintains consistent entity references across levels.

Step 2: Film KG Assembly. We assemble the structured descriptions, entity references, and explicitly identified relations from Step 1 into the Film KG, completing the Agentic Encoding:

$$G = E(V; P) = \mathrm{BuildGraph}\Big(\mathcal{C}_{\mathrm{film}}, \mathcal{B}_{\mathrm{reg}}, \{\mathcal{C}_{\mathrm{shot},i}, \{\mathcal{C}_{\mathrm{kf},i,j}\}_{j=1}^{J_i}\}_{i=1}^{S}\Big) \,. \tag{5}$$

$\mathrm{BuildGraph}$ deterministically maps the structured records to nodes and their identified relations and registry references to edges under the Film KG schema. See Secs. 4.2, E for details.

4.2 FILM KG REPRESENTATION

To preserve the cinematic multimodal dependencies discussed in Sec. 1 in an agent-native form, we organize the encoder's structured descriptions and identified relations into a Film KG: $G = (\mathcal{N}_G, \mathcal{E}_G, \mathcal{A}_G)$, where $\mathcal{N}_G$ contains structured-text nodes storing cinematic contents, $\mathcal{E}_G$ contains typed edges expressing content dependencies, and $\mathcal{A}_G$ stores linked generated multimodal assets.

In order to make multimodal content individually accessible and editable by agents, we organize reconstruction-critical descriptions into separately addressable textual nodes of types {Story, Event, Shot, Character, Scene, Object, Camera, Style, Audio, Keyframe}, while their edges to generated assets connect textual edits to multimodal creation.

Typed edges preserve structural, asset, temporal, and semantic dependencies among separately editable nodes, making their within- and cross-shot relationships explicit for coordinated editing.

Given edited nodes $Z \subseteq \mathcal{N}_G$ and edit type e, we determine the affected node set $\mathcal{N}_{\text{aff}}$ through edit-conditioned dependency traversal:

$$B = \mathrm{BFS}_e(G, Z), \qquad \mathcal{N}_{\text{aff}} = \mathrm{Up}\big(B \cup \mathrm{Check}_e^1(B)\big)\,. \tag{6}$$

Here, BFS_e collects the edited nodes and their propagation dependents into B; Check_e^1 adds nodes requiring one-hop consistency checks; and Up retains the selected nodes and adds their enclosing events and story so that higher-level descriptions reflect the edited shots. (See Sec. 5.4 and Appendix K, L.2.)

4.3 DUAL-LOOP TEXTUAL-GRADIENT EVOLUTION

To address Challenge 3 in Sec. 1, we develop dual-loop textual-gradient evolution for reliable agentic self-evolution. Extending the auto-encoding principle to agents, a fixed decoder exposes reconstruction errors, which are translated into textual gradients to revise either the shared encoding policy P or the input-specific Film KG G. Accordingly, the outer loop performs *Data-Agnostic Encoding Policy Pseudo-Training* (Sec. 4.3.2) to learn reusable encoding rules, while the inner loop performs *Data-Dependent KG Representation Refinement* (Sec. 4.3.3) to correct input-specific errors.

Textual-gradient updates. We adopt textual gradients (Yuksekgonul et al., 2024) to convert reconstruction errors into target-specific natural-language critiques for agents. For an update target $z \in \{P, G\}$ denoting encoding policy and KG in Figure 2, let $\mathcal{F}_z$ denote its relevant reconstruction errors. The agent generates a textual gradient and instantiates it as a challenger:

$$\mathbf{g}_{\text{text}}^z = \nabla_{\text{text}}(z, \mathcal{F}_z), \qquad z' = \mathrm{Propose}_z(z, \mathbf{g}_{\text{text}}^z), \qquad z \in \{P, G\}. \tag{7}$$

Here, ∇_{text} diagnoses how the current target should update, while $\mathrm{Propose}_z$ applies critiques to the encoding policy or selected generation-description nodes in G, with foundation-model weights fixed.

To determine whether a challenger should replace the current solution, we first define optimization signals and evaluation metrics for assessing reconstruction improvements and regressions (Sec. 4.3.1), then describe how each loop organizes updates and controls their acceptance (Secs. 4.3.2 and 4.3.3).

4.3.1 RECONSTRUCTION RESIDUAL

The reconstruction discrepancies $R(V, \hat{V})$ in Sec. 3.2 inform two fidelity measures, both higher-is-better: R_{reward} guides self-evolution, while R_{eval} is reserved for final evaluation to guard against reward exploitation.

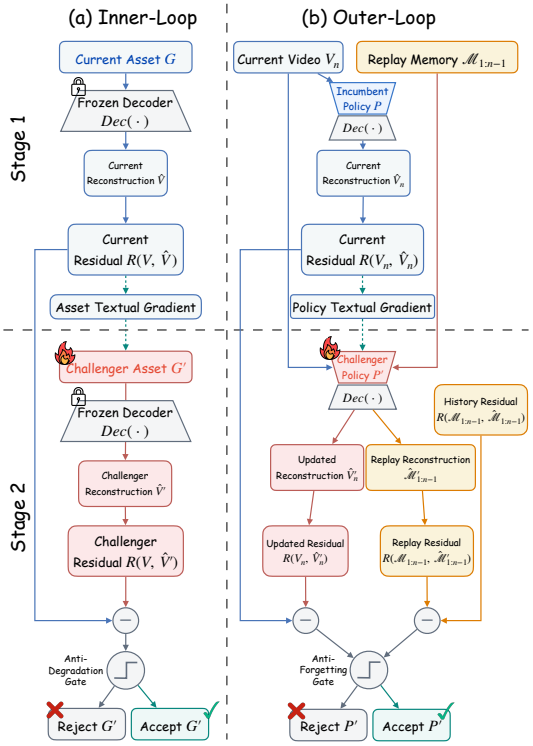
Optimization signal R_{reward}. Local revisions require fine-grained feedback, but holistic VLM scores can be coarse and unstable for dense multimodal content (Wang et al., 2026). We therefore decompose reconstruction quality into factual checks against the source at the decoder's keyframe and shot granularities. For each selected source keyframe or shot x, we construct $\mathcal{Q}(x) = \{(q_k, y_k)\}_{k=1}^K$: K binary questions q_k covering observable content attributes, paired with source-derived answers y_k, $k \in \{1, \ldots, K\}$. Both remain fixed across challenger comparisons to ensure a consistent factual reference. For reconstruction $\hat{x}$, the reward measures the fraction of checked source facts preserved:

$$R_{\text{reward}}^{\text{QA}}(x, \hat{x}) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}[\mathrm{Answer}(\hat{x}, q_k) = y_k]\,, \tag{8}$$

where Answer returns a binary answer from the reconstructed content and $\mathbb{I}[\cdot]$ is the indicator function. For these QA checks, the corresponding loss in Eq. 2 is $1 - R^{\mathrm{QA}}_{\mathrm{reward}}(x, \hat{x})$ (Appendix Sec. C).

While aggregated reward compares challengers, the failed checks identify discrepancies to guide revision. Letting $\hat{y}_k = \mathrm{Answer}(\hat{x}, q_k)$, we retain $\mathcal{F}_{\mathrm{QA}}(x, \hat{x}) = \{(q_k, y_k, \hat{y}_k) \mid 1 \le k \le K,\ \hat{y}_k \neq y_k\}$ as concrete evidence for textual-gradient updates in Eq. 7. Additional loop-specific diagnostics appear in Secs. 4.3.2 and 4.3.3; Appendix Sec. A details all diagnostic sources and their conversion into revision instructions.

Final evaluation R_{eval}. We assess final reconstruction fidelity with R_{eval} to guard against reward exploitation. It aggregates fine-grained factual judgments under a fixed protocol distinct from the optimization reward, with comparison directions and dimensions introduced in Sec. 5.1 and Appendix B.



Algorithm 1 Inner Loop: KG Representation Refinement

Input: Source video V, initial KG G, setting $\beta \in \{\mathrm{KF}, \mathrm{shot}\}$, budget T_{inner}
Output: Refined KG G

Prepare fixed verification checks from V; $\hat{V} \leftarrow \mathrm{Dec}(G)$
for $t = 1, \ldots, T_{\mathrm{inner}}$ **do**
  Diagnose reconstruction errors in $\hat{V}$ against V for setting β
  if no errors are found **then break**
  Select a keyframe or shot and target dimension $d^\star$; collect error evidence $\mathcal{F}_G$
  $G' \leftarrow \mathrm{Propose}_G(G, \nabla_{\mathrm{text}}(G, \mathcal{F}_G))$
    ▷ *Revise selected descriptions (Eq. 7)*
  $\hat{V}' \leftarrow \mathrm{Dec}(G')$; extract corresponding reconstructions u, u' from $\hat{V}, \hat{V}'$
  $\Gamma^{\beta}_{\mathrm{inner}}(u, u') \leftarrow \mathbb{I}\big[\mathrm{Improve}_\beta(u, u') \wedge \mathrm{Preserve}_\beta(u, u')\big]$
    ▷ *Anti-Degradation Gate*
  if $\Gamma^{\beta}_{\mathrm{inner}}(u, u') = 1$ **then**
    $(G, \hat{V}) \leftarrow (G', \hat{V}')$
  end if
end for
return G

Figure 2: Dual-loop textual-gradient evolution. (a) KG refinement with the Anti-Degradation Gate. (b) Encoding-policy pseudo-training with the Anti-Forgetting Gate.

4.3.2 Outer Loop: Encoding Policy Pseudo-Training

Before deployment, the outer loop learns a shared encoding policy from reconstruction feedback on a small video collection to improve fidelity on unseen inputs (Algorithm 2). Failed QA checks expose missing or distorted source facts and drive textual-gradient updates to P_b, $b \in \{\mathrm{shot}, \mathrm{kf}\}$, via Eq. 7 with $z = P$. Accepted revisions accumulate across videos into reusable general encoding rules.

Adapting the policy to the current video may weaken previously acquired capabilities. We therefore introduce the *Anti-Forgetting Gate*, requiring current-video reward improvement while limiting historical degradation (Algorithm 2, Line 13). Historical performance is compared against P_{ref}, the policy frozen at the start of the current video, to constrain cumulative regression across updates. Replay covers every preceding video while sampling assets and factual checks within each to limit verification cost. See Sec. 5.3 for gate ablations and Appendix Sec. D for sampling and thresholds.

In Algorithm 2, $\mathrm{Assess}(P, \mathcal{S}, \mathcal{Q})$ encodes and reconstructs samples $\mathcal{S}$ under P, returning mean QA reward and failed checks against $\mathcal{Q}$; $\mathrm{AssessHist}$ returns challenger and reference rewards on the same

Algorithm 2 Outer Loop: Encoding Policy Pseudo-Training

Input: Videos $\{V_i\}_{i=1}^{L}$, policy P_0, component b, budget T_{outer}, replay counts m, n
Output: Pseudo-trained component P_b^*

$P \leftarrow P_0$; replay memory $\mathcal{M} \leftarrow \varnothing$
for each V_i **do**
  $P_{\text{ref}} \leftarrow P$; prepare samples $\mathcal{S}_i$ and fixed checks $\mathcal{Q}_i$
  for $t = 1, \ldots, T_{\text{outer}}$ **do**
    $(R_c, \mathcal{F}_{\text{QA}}) \leftarrow \text{Assess}(P, \mathcal{S}_i, \mathcal{Q}_i)$
    $P' \leftarrow \text{Propose}_P(P, \nabla_{\text{text}}(P, \mathcal{F}_{\text{QA}}))$
      ▷ *Update only P_b (Eq. 7)*
    $(R'_c, _) \leftarrow \text{Assess}(P', \mathcal{S}_i, \mathcal{Q}_i)$
    $(R'_h, R_h^{\text{ref}}) \leftarrow (0, 0)$
    if $\mathcal{M} \neq \varnothing$ **then**
      $(\mathcal{S}_h, \mathcal{Q}_h) \leftarrow \text{Replay}(\mathcal{M}, m, n)$
      $(R'_h, R_h^{\text{ref}}) \leftarrow \text{AssessHist}(P', P_{\text{ref}}, \mathcal{S}_h, \mathcal{Q}_h)$
    end if
    $\Gamma_{\text{outer}} \leftarrow \mathbb{I}\big[(R'_c > R_c + \delta) \wedge (R'_h \geq R_h^{\text{ref}} - \epsilon_{\text{hist}})\big]$
    if $\Gamma_{\text{outer}} = 1$ **then** $P \leftarrow P'$
  end for
  Add $(\mathcal{S}_i, \mathcal{Q}_i)$ to $\mathcal{M}$
end for
return $P_b^* \leftarrow P_b$

Table 1: **Reconstruction fidelity** (%; higher is better). V/KF: direct video/keyframe comparison; V-BC/KF-BC: video/keyframe back-captioning. Overall averages the four scores.

Method	V	KF	V-BC	KF-BC	Overall
VideoAnalyzer	26.1	28.5	9.7	21.7	21.5
Storyboard Studio	16.4	28.6	9.6	23.0	19.4
soap2soap	36.7	39.5	15.8	21.3	28.3
AVA-Encoder (ours)	**57.8**	**73.7**	**29.7**	**34.6**	**49.0**

Table 2: **Reconstruction ablations** (%; higher is better). Except for the single-level encoder, all settings use hierarchical encoding. Single-loop ablations retain the other loop and its gate; gate ablations retain both loops.

Ablation setting	V	KF	V-BC	KF-BC	Overall
Single-level encoder	35.1	38.4	15.4	21.1	27.5
w/o both loops	50.7	67.6	21.2	29.9	42.4
Human-tuned policy	53.5	68.1	25.8	30.2	44.4
w/o policy pseudo-training	52.7	71.8	23.9	33.1	45.4
w/o KG refinement	55.6	68.3	26.6	32.7	45.8
w/o inner-loop gate	56.8	66.7	27.2	28.9	44.9
w/o outer-loop gate	52.9	70.2	29.8	33.0	46.5
w/o both gates	53.1	67.2	24.3	29.4	43.5
AVA-Encoder (full)	**57.8**	**73.7**	**29.7**	**34.6**	**49.0**

replay samples and checks. R_c and R_h denote current and historical rewards, with improvement margin δ and historical degradation tolerance ϵ_{hist}.

4.3.3 Inner Loop: KG Representation Refinement

A shared encoding policy, fixed at inference, may still overlook video-specific details. The inner loop addresses these omissions by autonomously refining the Film KG at test time while keeping P fixed (Algorithm 1). Reconstruction errors identify a keyframe or shot and a content dimension requiring correction. Setting $z = G$ in Eq. 7, the agent forms a textual gradient for this local discrepancy and revises the corresponding generation descriptions, which directly guide image or video reconstruction.

Improving one dimension may compromise content preserved in others. We therefore introduce the *Anti-Degradation Gate*, which accepts a challenger only if Improve_β confirms reconstruction improvement and Preserve_β limits degradation to the allowed tolerances for setting β (Algorithm 1, Line 8). Both checks use the same source content and fixed verification criteria. See Sec. 5.3 for refinement and gate ablations, and Appendix Sec. D for setting-specific checks and thresholds.

5 Experiments

We first evaluate video reconstruction fidelity across agentic systems, then isolate the contributions of hierarchical encoding, dual-loop optimization, and acceptance gates through ablation, and finally assess Film KG representation's utility for coordinated editing and downstream generation.

5.1 Experimental Setup

Data. We pseudo-train on 6 videos and evaluate on 18 non-overlapping videos spanning animation, human-directed AI short films, and classic cinema, detailed in Appendix Sec. H .

Models and comparison protocol. Qwen-3.7-Plus (Alibaba Cloud, 2026b) performs hierarchical understanding within the Agentic Video Encoder and revises the policy or Film KG using textual

gradients; Gemini-3.1-Pro-Preview (Google AI, 2026a) performs factual judgments for R_{reward} and R_{eval}. Nano Banana Pro (Google AI, 2026b) and HappyHorse 1.0 (Alibaba Cloud, 2026a) generate decoder images and videos, respectively. All model weights remain fixed. VideoAnalyzer (Docusphere, 2026), Storyboard Studio (BroderQi, 2026), soap2soap (Song et al., 2026), and AVA-Encoder share the same understanding model, test videos, generators, and per-shot decoder-interface limits of 5 generated reference keyframes and 1,200 shot-prompt tokens. Reconstruction generators never receive source-video pixels. Appendix Secs. N and G provide prompts and baseline adaptation protocols.

Evaluation. We assess retained information at two granularities, videos and keyframes, through both direct comparison and comparison of independently generated descriptions, yielding four evaluation directions: V directly compares audiovisual and temporal content; KF examines static visual details. Since all compared methods use text-centered representations, V-BC and KF-BC compare detailed semantic facts in independently generated source and reconstruction descriptions. These complementary granularities and comparison modes reduce reliance on any single view.

Each direction examines eight dimensions: Character, Scene, Position, Motion, Audio, Style, Camera, and Narrative. This decomposition turns coupled cinematic information into fine-grained factual judgments, which fixed rules aggregate into direction scores. Keyframe directions exclude Audio, and Overall is the mean of the four direction scores. Final-evaluator (R_{eval}) rankings agree with human preferences in 97.3% of 730 blinded pairwise comparisons, detailed in Appendix Sec. B.

5.2 Reconstruction Fidelity

Under the shared decoder and interface budget, AVA-Encoder reaches 49.0% Overall, surpassing the strongest baseline, soap2soap, by 20.7 percentage points and leading in all four directions (Table 1). Gains across direct and caption-based comparisons indicate better preservation of perceptual detail and semantic information. Figure 3 shows a representative comparison; Appendix Secs. I and L provide dimension-level results and further examples.



Figure 3: Reconstruction comparison between AVA-Encoder and baselines, showing the preservation of character identity, visual detail, and visual consistency.

5.3 Ablation Study

Table 2 examines encoding structure, policy choice, dual-loop optimization, and acceptance gates under the same reconstruction evaluation protocol.

First, hierarchical encoding raises Overall from the single-level encoder's 27.5% to 42.4% with both loops disabled, supporting multi-level understanding for faithful representation. Second, the outer loop learns a shared encoding policy that improves reconstruction on unseen videos: pseudo-training on six videos raises Overall on 18 unseen videos by 3.4 points without the inner loop (42.4% to 45.8%) and 3.6 with it (45.4% to 49.0%). In the policy-only setting, the evolved policy also surpasses the human-tuned policy (44.4%) by 1.4 points, while using 74.3% fewer shot-level and 70.1% fewer keyframe-level system-prompt tokens. Thus, policy evolution improves fidelity over human tuning

Table 3: **Downstream video-generation quality** (1–4; higher is better). Each cell reports scores without/with a single textual injection of the complete Film KG. Bold indicates an improvement; AV denotes audiovisual quality.

Method	Character	Plot	Camera	Style	AV	**Overall**
MovieAgent (Wu et al., 2025)	3.27/**3.57**	1.55/**3.33**	3.35/**3.40**	2.23/**2.75**	1.58/**2.18**	2.39/**3.04**
FilmAgent (Xu et al., 2025)	1.93/**3.23**	1.58/**2.08**	3.48/**3.63**	1.60/**3.13**	1.30/**1.45**	1.98/**2.70**
Anim-Director (Li et al., 2024)	1.43/**3.33**	2.68/1.55	3.20/**3.63**	1.45/**2.48**	1.30/**1.55**	2.01/**2.51**
VideoStudio (Long et al., 2024)	3.00/1.43	1.28/**1.60**	3.03/**3.40**	1.48/**2.70**	1.33/**1.70**	2.02/**2.17**

while producing more compact prompts. Third, the inner loop addresses input-specific errors, raising Overall by 3.2 points with the pseudo-trained policy fixed (45.8% to 49.0%) and 3.0 with the initial policy (42.4% to 45.4%). This refinement complements shared-policy optimization: together, the loops gain 6.6 points over the initial configuration. Finally, both acceptance gates improve fidelity: removing only the inner-loop Anti-Degradation Gate or outer-loop Anti-Forgetting Gate lowers Overall from 49.0% to 44.9% or 46.5%, respectively; removing both yields 43.5%. These ablations support acceptance control in both KG refinement and policy learning. Appendix Sec. J provides configuration definitions and supplementary analyses.

5.4 Representation Applications

Graph-based editing. Film KG supports targeted edits with coordinated updates to dependent content. Figure 4 shows consistent character replacement across six linked shots, preserving unrelated characters, scenes, and event structure. Appendix Sec. L.2 provides further edit examples.



Figure 4: Editing Showcase: Character replacement across a six-shot sequence. A single character edit jointly updates the dependent keyframes and shots.

Downstream video generation. We provide each video-creation framework with the complete Film KG as text and a one-sentence usage request, without workflow changes, framework-specific adapters, or prompt tuning. Overall improves across all tested frameworks (Table 3), demonstrating cross-system reuse via a simple textual interface. Evaluation details appear in Appendix Sec. M.

6 Conclusion

We present AVA-Encoder, an agentic video auto-encoding framework that transforms cinematic videos into Film KG representation. Reconstruction-driven dual-loop self-evolution improves shared encoding policies and video-specific representations without ground-truth graphs or expert creation trajectories. AVA-Encoder achieves 49.0% Overall reconstruction fidelity, a 20.7-percentage-point gain over the strongest baseline, while enabling coordinated editing and improving downstream video generation. These results demonstrate a path toward bridging video and agent spaces, making video creation knowledge in finished films explicit, editable, and reusable by agents.

AI USE STATEMENT

Among the tasks requiring disclosure under the ICLR 2027 AI Policy for Authors, we used generative AI tools only to assist with translation; we did not use them for any other such task. Additionally, we used generative AI tools to edit the manuscript for clarity, concision, readability, and presentation, including minor LaTeX and formatting adjustments. Separately, foundation models are used as components of our proposed system and experimental pipeline; their roles and configurations are described in Sec. 5.1, with implementation details and prompts provided in Appendix Sec. E, Appendix Sec. F, and Appendix Sec. N. All AI-assisted manuscript edits were reviewed by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

Our work uses open-sourced video content for research on video representation, reconstruction, and editing. We do not claim ownership of the underlying source media, and the Film KG dataset contains only derived structured-text representations rather than the original image, audio, or video assets. The demonstrated editing capabilities, including identity replacement, are intended to study controllable and dependency-aware video editing, but could potentially be misused for deceptive or unauthorized content manipulation; responsible use should respect applicable rights, consent, and legal requirements. We encourage users of the released representations and framework to respect and follow the licenses and rights associated with the underlying source content.

REPRODUCIBILITY STATEMENT

We provide the experimental setup, model configurations, comparison protocol, and representation budget in Sec. 5.1. Additional details required to reproduce the method and experiments are provided throughout the appendix: Appendix Sec. A describes the textual-gradient formulation and self-optimization procedure; Appendix Sec. B specifies the reconstruction evaluation metrics and aggregation protocol; Appendix Sec. C defines the optimization reward used by the dual loops; Appendix Sec. D specifies the dual-loop acceptance gates and pseudo-training procedure; Appendix Sec. E details the Agentic Video Encoder implementation; Appendix Sec. F provides its system prompts; Appendix Sec. G documents baseline adaptation and the shared comparison protocol; Appendix Sec. H specifies the released dataset, pseudo-training data, and evaluation split; Appendix Sec. J defines the ablation settings; Appendix Sec. K details the Film KG schema and editing operations; Appendix Sec. M specifies the downstream story-video evaluation; and Appendix Sec. N provides the complete prompts used in our experiments.

REFERENCES

Alibaba Cloud. HappyHorse 1.0 reference-to-video model. Alibaba Cloud Model Studio Documentation, 2026a. URL https://www.alibabacloud.com/help/en/model-studio/video-generate-edit-model.

Alibaba Cloud. Qwen3.7-Plus Model Information. Alibaba Cloud Model Studio Documentation, 2026b. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-7-plus.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Elias Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. Goldfish: Vision-language understanding of arbitrarily long videos. In *Computer Vision – ECCV 2024*, pp. 251–267. Springer, 2024.

BroderQi. Storyboard studio: Ai-powered professional storyboarding workbench. GitHub repository, 2026. URL https://github.com/BroderQi/Storyboard.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI Technical Report, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/. OpenAI Sora.

ByteDance Seed Team. Seedance 2.0. Official Model Page, 2026. URL https://seed.bytedance.com/seedance2_0.

Meng Chu, Yicong Li, and Tat-Seng Chua. GraphVideoAgent: Enhancing long-form video understanding with entity relation graphs. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4639–4648. Association for Computing Machinery, 2025. doi: 10.1145/3746027.3755537. URL https://doi.org/10.1145/3746027.3755537.

Jiwan Chung and Youngjae Yu. Long story short: A summarize-then-search method for long video question answering. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2023.

comfyanonymous and ComfyUI contributors. ComfyUI: The most powerful and modular diffusion model gui and backend. https://github.com/comfyanonymous/ComfyUI, 2023. Accessed: 2026-06-29.

Docusphere. video-analyzer: Agentic video transcript generation. GitHub repository, 2026. URL https://github.com/docusphere/video-analyzer.

Google AI. Gemini 3.1 Pro Preview. Google AI for Developers Documentation, 2026a. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview.

Google AI. Gemini 3 Pro Image (Nano Banana Pro). Google AI for Developers Documentation, 2026b. URL https://ai.google.dev/gemini-api/docs/models/gemini-3-pro-image.

Google DeepMind. Veo: Our state-of-the-art video generation model. Google DeepMind Technical Report, 2024. URL https://deepmind.google/technologies/veo/.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13504–13514, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. MovieNet: A holistic dataset for movie understanding. In *Computer Vision – ECCV 2020*, pp. 709–727. Springer, 2020.

Kuaishou Technology. Kling: Kuaishou text-to-video generation model. Kuaishou Technical Report, 2024. URL https://kling.kuaishou.com/.

Jinxiang Lai, Zexin Lu, Jiajun He, Rongwei Quan, Wenzhe Zhao, Qinyu Yang, Qi Chen, Qin Lin, Chuyue Li, Tao Gao, Yuhao Shan, Shuai Shao, Song Guo, and Qinglin Lu. VisionCreator: A native visual-generation agentic model with understanding, thinking, planning and creation, 2026.

Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Jiashun Zhu, Jinyi Xu, Zhen Zhao, and Min Zhang. Anim-Director: A large multimodal model powered agent for controllable animation video generation. In *SIGGRAPH Asia 2024 Conference Papers*. ACM, 2024. doi: 10.1145/3680528.3687688.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideoDirectorGPT: Consistent multi-scene video generation via LLM-guided planning. In *Conference on Language Modeling (COLM)*, 2024.

Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. VideoStudio: Generating consistent-content and multi-scene videos. In *European Conference on Computer Vision (ECCV)*. Springer, 2024. doi: 10.1007/978-3-031-73027-6_27.

Louis Mahon and Mirella Lapata. ScreenWriter: Automatic screenplay generation and movie summarisation, 2024.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, et al. Movie Gen: A cast of media foundation models, 2024. URL https://arxiv.org/abs/2410.13720.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. Agentic very long video understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46575–46602. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2161. URL https://aclanthology.org/2026.acl-long.2161/.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232, 2024.

Yiren Song, Huilin Zhong, Kevin Qinghong Lin, Haofan Wang, and Mike Zheng Shou. Soap2Soap: Long cinematic video remaking via multi-agent collaboration, 2026. URL https://arxiv.org/abs/2605.17423.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Haozhe Wang, Long Li, Chao Qu, Weidi Xu, Fengming Zhu, Wei Chu, and Fangzhen Lin. To code or not to code? adaptive tool integration for math language models via expectation-maximization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3060–3075, 2025.

Haozhe Wang, Qixin Xu, Changpeng Wang, Taofeng Xue, Chong Peng, Wenhu Chen, and Fangzhen Lin. Bad seeing or bad thinking? rewarding perception for multimodal reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent CoT planning, 2025.

Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawendé F. Bissyandé, and Saad Ezzini. DreamFactory: Pioneering multi-scene long video generation with a multi-agent framework, 2024.

Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, Jun Yu, and Min Zhang. FilmAgent: A multi-agent framework for end-to-end film automation in virtual 3D spaces, 2025.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. MAGVIT: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Cai, Bin Lin, Li Yuan, Lifang He, Chi Wang, Yanfang Ye, and Lichao Sun. Mora: Enabling generalist video generation via a multi-agent framework, 2024.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic "differentiation" via text, 2024. URL https://arxiv.org/abs/2406.07496.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pp. 543–553, 2023.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision, 2024a. LongVA.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data, 2024b.

Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8806–8817, 2024.

Appendix

Organization and relation to the main paper. The main paper presents the task formulation (Sec. 3), the AVA-Encoder method (Sec. 4), and the experiments (Sec. 5). This appendix follows the same order: it first expands the optimization signals and loops of Sec. 4.3, then the encoder of Sec. 4.1, then the experimental protocol and results of Sec. 5, and finally the Film KG of Sec. 4.2 and its applications.

Optimization and evaluation.

- Section A expands the textual-gradient update in Eq. 7 by defining the correction records behind $\mathcal{F}_G$ and $\mathcal{F}_P$.
- Section B defines the final evaluation R_{eval}: its four directions, eight dimensions, fact-level scoring, aggregation, and agreement with human judgments.
- Section C defines the optimization reward R_{reward} and its instances used by Algorithms 2 and 1.
- Section D instantiates the Anti-Degradation and Anti-Forgetting Gates and the sequential pseudo-training procedure.
- Section E details preprocessing, hierarchical understanding, and deterministic Film KG assembly for the Agentic Video Encoder.
- Section F identifies the initial, pseudo-trained, and human-tuned encoding policies compared in the Ablation Study (Sec. 5.3).

Experimental setup and numerical results.

- Section G describes baseline adaptation, the shared decoder interface, and source-pixel exclusion (Sec. 5.1).
- Section H describes the released Film KG dataset and the pseudo-training and evaluation videos.
- Section I reports per-dimension V and KF scores behind Table 1 (Reconstruction Fidelity, Sec. 5.2).
- Section J defines every configuration in Table 2 (Ablation Study, Sec. 5.3) and reports effect sizes.

Film KG and qualitative evidence.

- Section K specifies the stored Film KG schema and its construction from encoder records (Sec. 4.2).
- Section L provides further reconstruction comparisons and graph-based editing examples, including the implementation of the edit scope in Eq. 6.
- Section M specifies the evaluation for downstream video generation (Sec. 5.4).

Complete experimental specification.

- Section N reproduces complete English translations of the Chinese system prompts used for the Agentic Video Encoder, R_{reward}, and R_{eval}.

A Textual Gradients and AVA-Encoder Self-Optimization

This section expands the textual-gradient update in main-paper Sec. 4.3 (Eq. 7). It first recalls the TextGrad abstraction, then specifies where the reconstruction errors $\mathcal{F}_z$ come from in each optimization setting and how they are turned into revision instructions.

Table 4: Sources of the reconstruction errors $\mathcal{F}_z$ in Eq. 7. Each row specifies which grounded reconstruction failure supplies a correction record.

Optimization setting	Reconstruction feedback	Fact selected for correction
KG refinement: keyframe	Direct GT–current-keyframe difference diagnosis	A grounded difference that specifies its dimension and image location together with the GT and current-reconstruction observations.
KG refinement: video shot	Fixed shot-reward checklist	A GT-derived reward item judged as mismatch or absent in the reconstructed shot.
Policy pseudo-training: P_{shot} branch	Frozen video QA banks over the current video's selected shots	A question whose answer on a reconstructed shot differs from its GT-derived expected answer.
Policy pseudo-training: P_{kf} branch	Frozen image QA banks over the current video's selected keyframes	A question whose answer on a reconstructed keyframe differs from its GT-derived expected answer.

A.1 TextGrad Background

TextGrad represents an LLM workflow as a directed computation graph whose variables may contain text (Yuksekgonul et al., 2024). Its evaluation feedback is written in natural language and passed backward to the text variable that should be revised. For a text variable z, the relation between evaluation, textual gradient, and revision is summarized by

$$\mathbf{g}_{\text{text}}(z) := \nabla_{\text{text}}\mathcal{L}_{\text{TG}}(z), \qquad z^{+} := \text{UpdateText}(z, \mathbf{g}_{\text{text}}(z)). \tag{9}$$

Here, $\mathcal{L}_{\text{TG}}$ is the illustrative TextGrad evaluation objective, ∇_{text} denotes backward assignment of natural-language feedback, and $\mathbf{g}_{\text{text}}(z)$ is the resulting textual gradient for z. UpdateText revises z according to that feedback, producing z^{+}. The textual gradient is not a numerical derivative; it identifies the error caused by z and states how z should change. AVA-Encoder instantiates this abstraction with reconstruction errors as feedback, $\nabla_{\text{text}}(z, \mathcal{F}_z)$ as the feedback operator, and Propose_z as the update (Eq. 7).

A.2 Structured Textual Gradients in AVA-Encoder

Sources of reconstruction errors. AVA-Encoder derives textual gradients from differences between the source video and its reconstruction. Throughout the appendix, ground truth (GT) refers to this source content (a source video, shot, or keyframe), not to a ground-truth Film KG. Table 4 lists the source of $\mathcal{F}_z$ in each optimization setting; it expands the error diagnosis in Algorithm 1 ($z = G$) and the failed checks $\mathcal{F}_{\text{QA}}$ in Algorithm 2 ($z = P$). For KG refinement, the keyframe setting $\beta = \text{KF}$ directly diagnoses grounded differences between the GT and reconstructed keyframes, whereas the shot setting $\beta = \text{shot}$ uses failed items of a fixed shot-reward checklist. For policy pseudo-training, each prompt-rewriting branch uses the failed QA checks of its own granularity: video QA for P_{shot} and image QA for P_{kf}. In all cases, one grounded difference or one failed check is a single fact-level reconstruction failure, the basic unit from which correction records are built.

Correction records. Regardless of its source, we represent each selected failure $\xi \in \mathcal{F}_z$ as the following correction record:

$$\mathbf{a} := \text{Corr}(\xi) = \left(d, u^{\text{GT}}, u^{\text{rec}}, e, h\right). \tag{10}$$

Here, d is the evaluation dimension of the failure, u^{GT} is the GT-derived fact, u^{rec} is the corresponding fact observed in the reconstruction, e is the supporting visual or audio evidence, and h is the requested correction.

From records to updates. Let $\mathcal{T}_z := \{\text{Corr}(\xi) : \xi \in \mathcal{F}_z\}$ be the structured record set for update target $z \in \{G, P\}$. It is an intermediate representation of $\mathcal{F}_z$, so the operators of Eq. 7 are implemented as

$$\begin{aligned} \mathbf{g}^{G}_{\text{text}} &= \nabla_{\text{text}}(G, \mathcal{F}_G) \text{ computed from } \mathcal{T}_G, & G' &= \text{Propose}_G(G, \mathbf{g}^{G}_{\text{text}}), \\ \mathbf{g}^{P}_{\text{text}} &= \nabla_{\text{text}}(P, \mathcal{F}_P) \text{ computed from } \mathcal{T}_P, & P' &= \text{Propose}_P(P, \mathbf{g}^{P}_{\text{text}}). \end{aligned} \tag{11}$$

$\mathbf{g}^{G}_{\text{text}}$ is an asset-level revision instruction for the generation descriptions of the selected keyframe or shot, and $\mathbf{g}^{P}_{\text{text}}$ is a policy-level revision instruction for the encoding prompts. For the outer loop, $\mathcal{F}_P = \mathcal{F}_{\text{QA}}$ collected over the current video. The policy update runs in two separate prompt-rewriting branches: one revises P_{shot} with P_{kf} fixed, and the other revises P_{kf} with P_{shot} fixed; both keep P_{film} fixed. Every challenger G' or P' is evaluated by the corresponding gate before acceptance (Sec. D).

B Reconstruction Evaluation Metrics

This section specifies the final evaluation R_{eval} used in the Experimental Setup (Sec. 5.1) and Reconstruction Fidelity (Sec. 5.2). The main paper evaluates reconstruction along four comparison directions—Video (V), Keyframe (KF), Video back-captioning (V-BC), and Keyframe back-captioning (KF-BC)—each decomposed into a shared set of cinematic dimensions. We define the directions and dimensions (Sec. B.1), score individual facts (Secs. B.2–B.3), aggregate them into the reported scores (Sec. B.5), and validate the result against human judgments (Sec. B.6).

R_{eval} uses dedicated evaluation prompts. Under these fixed prompts, the evaluator constructs or extracts atomic source-grounded facts and outputs a categorical judgment with supporting evidence for each fact; fixed deterministic rules then map the categories to numerical values (Eqs. 14 and 15) and aggregate them into the reported scores (Eq. 16). This decomposition avoids asking the model for one direct score on a complex image or video.

B.1 Reconstruction Directions and Evaluation Dimensions

We evaluate every representation system from four directions: direct Video (V), direct Keyframe (KF), Video back-captioning (V-BC), and Keyframe back-captioning (KF-BC). Let

$$\begin{aligned} \mathcal{D}_{\text{dir}} &:= \{\text{V}, \text{KF}, \text{V-BC}, \text{KF-BC}\}, \\ \mathcal{D}_{\text{eval}} &:= \{\text{Character}, \text{Scene}, \text{Position}, \text{Motion}, \\ &\qquad \text{Audio}, \text{Style}, \text{Camera}, \text{Narrative}\}. \end{aligned} \tag{12}$$

The four directions examine different types of evidence. A high-quality agentic video representation should retain the audiovisual and temporal information needed to reconstruct the complete source video, which motivates direct V evaluation. Because dense multimodal video makes small static visual details difficult to inspect fully, direct KF evaluation isolates identity, object, layout, composition, lighting, and texture evidence. AVA-Encoder and all compared baselines also use text-centered structured representations; V-BC and KF-BC therefore compare recoverable atomic facts in a common text space rather than requiring their internal structures to match. Together, the four directions measure end-to-end reconstruction, fine-grained visual fidelity, and text-level meaning consistency.

$\mathcal{D}_{\text{dir}}$ is the set of reconstruction directions reported in Table 1 of the main paper, and $\mathcal{D}_{\text{eval}}$ is the shared set of dimensions used by R_{eval}. We write $\mathcal{D}_{\text{eval}}$ to distinguish it from the policy-learning video collection $\mathcal{D}$ in main-paper Eq. 3. The directions combine two media—video shots and keyframes—with two observation spaces—direct comparison and blind back-captioning:

- **Video (V)** directly compares a source video shot with its reconstructed shot. The evaluator constructs source-grounded visual and audio checks and verifies each check against the reconstruction.
- **Keyframe (KF)** directly compares a source keyframe with its reconstructed keyframe. We use **KF** for keyframe throughout the remainder of the appendix; only evidence observable in a static image is evaluated.
- **Video back-captioning (V-BC)** describes the source and reconstructed video shots independently and compares atomic facts extracted from the two descriptions.

Table 5: Evaluation dimensions shared by the four reconstruction directions. Video directions use all eight dimensions; keyframe directions use seven because Audio is N/A.

Dimension	Evaluated attributes
Character	Face and identity, body/build, clothing, expression, gaze, and pose.
Scene	Scene type, background and set dressing, spatial layout, and environmental elements.
Position	Absolute character positions, inter-character spatial relations, and foreground/background depth.
Motion	Actions, interactions, motion direction, and speed; only static action/pose evidence is used for KF.
Audio	Speaker–dialogue–voice association, background music, and sound effects; N/A for KF and KF-BC.
Style	Artistic treatment, texture, color and temperature, illumination, contrast, and light direction.
Camera	Shot language, camera direction/speed, viewpoint, composition, and shot scale; only static attributes are used for KF.
Narrative	Visible events, event ordering, and emotional tone.

- **Keyframe back-captioning (KF-BC)** applies the same independent description and atomic-fact comparison to the source and reconstructed keyframes.

In the two back-captioning directions, each captioner observes only one side of the comparison. This separation prevents source information from entering the reconstruction description before atomic-fact scoring.

Video directions use all dimensions in $\mathcal{D}_{\text{eval}}$. A keyframe has no audio, so KF and KF-BC use $\mathcal{D}_{\text{eval}} \setminus \{\text{Audio}\}$, mark Audio as N/A, and average seven dimensions. Motion and Camera in the keyframe directions retain only attributes observable in a static image; cross-frame speed and camera movement are not scored.

B.2 Direct V and KF Scoring

For each applicable dimension $d \in \mathcal{D}_{\text{eval}}$, the evaluator first sees only the GT material and produces a checklist C_d of independently verifiable atomic facts. Here, a checklist is simply the set of concrete GT facts that must be checked for one dimension, such as a character identity, an object position, or a dialogue line. Critical identity, central action, scene, style, and dialogue facts form the subset C_d^{crit}. The evaluator then examines the reconstruction and assigns every fact $c \in C_d$ a verdict with visible or audible evidence. The numerical score for one fact is

$$w(c) = \begin{cases} 1, & \text{match,} \\ 0.5, & \text{partial,} \\ 0, & \text{mismatch or absent.} \end{cases} \tag{13}$$

Here, $w(c)$ is the fidelity score assigned to checklist item c: a complete match receives 1, a partial match receives 0.5, and an incorrect or missing fact receives 0.

The dimension score first averages the fact scores and then applies a cap when a critical fact is incorrect or missing:

$$\begin{aligned} \phi_d^{\text{raw}} &= \frac{1}{|C_d|} \sum_{c \in C_d} w(c), \\ \phi_d &= \begin{cases} \min(\phi_d^{\text{raw}}, 0.4), & \exists c \in C_d^{\text{crit}} : \ w(c) = 0, \\ \phi_d^{\text{raw}}, & \text{otherwise.} \end{cases} \end{aligned} \tag{14}$$

C_d is the checklist for dimension d, $|C_d|$ is its number of items, and $C_d^{\text{crit}} \subseteq C_d$ is the subset marked critical. ϕ_d^{raw} is the item average and ϕ_d is the reported dimension-fidelity score. The cap prevents a collection of minor matches from compensating when a reconstruction-critical fact is not satisfied.

B.3 Blind Back-Captioning Scoring

For V-BC and KF-BC, source and reconstruction are independently described for each evaluated attribute in Table 5 using concrete, verifiable statements. Atomic facts are extracted from the source description, and the reconstruction description labels each fact as a hit, conflict, or missing. The back-captioning score is the matched-fact rate with an additional penalty for contradictions:

$$\phi_{\text{fact}} = \text{clip}\left(\frac{n_{\text{hit}} - 0.5n_{\text{conflict}}}{n_{\text{facts}}}, 0, 1\right). \tag{15}$$

n_{facts} is the number of atomic facts extracted from the source description, n_{hit} counts facts reproduced by the reconstruction, and n_{conflict} counts contradicted facts. $\text{clip}(x, 0, 1)$ limits x to the interval $[0, 1]$; facts not mentioned by the reconstruction contribute neither a match nor a contradiction. Thus, a contradiction is penalized more strongly than an omission. KF-BC again uses seven dimensions and reports Audio as N/A.

B.4 System Prompts for Reconstruction Evaluation

The frozen R_{eval} prompts used in the reported experiments were written in Chinese; complete direct English translations are reproduced in Sec. N. Every compared representation system uses the same prompts.

B.5 Reconstruction-Benchmark Aggregation

This subsection explains how the fine-grained judgments produced by R_{eval} are converted step by step into the V, KF, V-BC, KF-BC, and Overall scores reported in Table 1 of the main paper. The averaging proceeds from facts to dimensions, from dimensions to shots or keyframes, from these units to video cases, and finally from case-level results to the four reconstruction directions and their Overall score.

The reconstruction benchmark compares four representation systems: AVA-Encoder, VideoAnalyzer, Storyboard Studio, and soap2soap. Let a index a representation system and let $u \in \mathcal{D}_{\text{dir}}$ index a reconstruction direction defined in Eq. 12. All four systems receive the same source videos and use the same frozen generation backbones.

Scores are averaged in the order fact, dimension, shot, case, and direction. For system a, video case v, shot s, and direction u, let $\mathcal{D}_{a,v,s,u} \subseteq \mathcal{D}_{\text{eval}}$ contain the applicable dimensions. Let $\psi_{a,v,s,u,d}$ denote the dimension score: it is ϕ_d from Eq. 14 for V and KF, and ϕ_{fact} from Eq. 15 for V-BC and KF-BC. The remaining averages are

$$\begin{aligned}
B_{a,v,s,u} &:= \frac{1}{|\mathcal{D}_{a,v,s,u}|} \sum_{d \in \mathcal{D}_{a,v,s,u}} \psi_{a,v,s,u,d}, \\
B_{a,v,u} &:= \frac{1}{|\mathcal{S}_{a,v,u}|} \sum_{s \in \mathcal{S}_{a,v,u}} B_{a,v,s,u}, \\
B_{a,u} &:= \frac{1}{|\mathcal{X}_{a,u}|} \sum_{v \in \mathcal{X}_{a,u}} B_{a,v,u}, \\
B_a^{\text{overall}} &:= \frac{1}{|\mathcal{D}_{\text{dir}}|} \sum_{u \in \mathcal{D}_{\text{dir}}} B_{a,u}.
\end{aligned} \tag{16}$$

$\mathcal{S}_{a,v,u}$ is the set of evaluated shots or selected keyframes for system a in video case v and direction u, and $\mathcal{X}_{a,u}$ is its set of evaluated video cases. $B_{a,v,s,u}$, $B_{a,v,u}$, and $B_{a,u}$ are the shot-, case-, and direction-level scores, respectively; B_a^{overall} is the unweighted mean of the four direction scores reported as Overall in Table 1 of the main paper.

Table 6: Two methodologically distinct signals built from the reconstruction residual: R_{reward} is used for optimization, whereas R_{eval} is used for final reporting.

	Final evaluation R_{eval}	Optimization reward R_{reward}
Purpose	Final comparison and per-dimension analysis	Diagnosis and candidate acceptance during self-optimization
Atomic judgment units	Source-grounded facts in each direction and dimension	Frozen QA facts for the outer loop; frozen QA facts and fixed-prompt fidelity scores for keyframe refinement; fixed reward-checklist facts for shot refinement
Primary benefit	Broad, common benchmark coverage	Localized correction evidence and comparison across rounds
Use	Reported tables and per-case analysis	Diagnosis and candidate selection in both loops

B.6 Alignment with Human Evaluation

This subsection defines the blinded pairwise study behind the 97.3% agreement reported in the Experimental Setup (Sec. 5.1).

We randomly sampled 10 videos and ran the self-optimization loops on them; the reconstructions produced along these runs yield 730 blinded triples, each containing a GT keyframe or shot and two generated candidates whose identities are hidden. Two expert annotators participated in the evaluation, judging which candidate is closer to GT. The automatic evaluator agrees with the human judgments on 710 of the 730 triples, yielding **97.3% agreement** $(710/730)$.

C Reconstruction Reward for Loop Optimization

This section expands the optimization reward R_{reward} of main-paper Sec. 4.3.1. We first contrast it with R_{eval}, then define the atomic QA reward, and finally show how its instances supply the errors and scores used by Algorithms 2 and 1.

Optimization reward versus final evaluation. Both R_{eval} and R_{reward} are derived from differences between a source and its reconstruction, but they serve different purposes (Table 6). R_{eval} is the shared four-direction benchmark protocol and is used only for final cross-system reporting. R_{reward} is the signal used inside the loops: its failed items diagnose the current solution, and its scalar value verifies challengers. The outer loop and keyframe refinement use frozen QA banks, while video-shot refinement uses a fixed shot-reward checklist (Sec. D). The two signals differ in their prompts, judgment units, and scoring rules.

C.1 Atomic QA Reward

For the QA-based instances, each selected source keyframe or shot x is decomposed into approximately 30 binary questions, each testing one observable fact. This avoids relying on one direct score for a complex image or video and gives the loop both a more precise reward and a specific error to correct. The bank $\mathcal{Q}(x) = \{(q_k, y_k)\}_{k=1}^{K}$ is generated once from the source and then frozen. Every current and challenger reconstruction $\hat{x}$ is evaluated against the same questions, so question-set changes never enter a reward difference. Restating main-paper Eq. 8 together with its failed checks,

$$\begin{aligned} \chi_k(\hat{x}) &:= \mathbb{I}[\text{Answer}(\hat{x}, q_k) = y_k], \qquad R^{\text{QA}}_{\text{reward}}(x, \hat{x}) := \frac{1}{K}\sum_{k=1}^{K} \chi_k(\hat{x}), \\ \mathcal{F}_{\text{QA}}(x, \hat{x}) &:= \{(q_k, y_k, \hat{y}_k) : \hat{y}_k \neq y_k\}, \qquad \hat{y}_k := \text{Answer}(\hat{x}, q_k). \end{aligned} \tag{17}$$

$\chi_k(\hat{x})$ equals one when the reconstruction gives the source-derived answer, so $R^{\text{QA}}_{\text{reward}}$ is a higher-is-better pass rate and $1 - R^{\text{QA}}_{\text{reward}}$ is the corresponding loss. Video QA spans the eight reconstruction

dimensions; image QA covers Character, Scene, and Composition facts observable in a static image. Each failed tuple $\xi = (q_k, y_k, \hat{y}_k) \in \mathcal{F}_{\mathrm{QA}}$ becomes a correction record (Eq. 10): the question's dimension supplies d, its expected answer and underlying source fact supply u^{GT}, the reconstruction answer supplies u^{rec}, the examined evidence supplies e, and the required factual change supplies h.

C.2 System Prompts for R_{reward}

Complete direct English translations of the Chinese prompts used to construct and answer the frozen video and image QA banks are reproduced in Sec. N.

C.3 R_{reward} Instances in Dual-Loop Optimization

Each loop setting uses one explicit instance of R_{reward} for challenger verification and one error source for its textual gradient (Table 4).

Inner loop, keyframe setting. For a GT keyframe I_{GT} and a reconstructed keyframe I, the verification score is a composite keyframe fidelity

$$R^{\mathrm{KF}}_{\mathrm{reward}}(I) := q_{\mathrm{char}}(I) \cdot q_{\mathrm{scene}}(I) \cdot \left(s_{\mathrm{style}}(I)\, s_{\mathrm{comp}}(I)\, s_{\mathrm{cl}}(I)\right)^{1/3}, \tag{18}$$

where q_{char} and q_{scene} are the pass rates of the character and scene questions in the frozen image QA bank (Eq. 17), and s_{style}, s_{comp}, and $s_{\mathrm{cl}} \in [0, 1]$ are fixed-prompt VLM scores of style, composition, and color-and-lighting fidelity to I_{GT}. The multiplicative form prevents a collapse in one aspect from being averaged away by the others. The textual gradient, however, is formed from direct image differences rather than failed checks:

$$\mathcal{F}^{\mathrm{KF}}_{G} := \mathrm{Diff}_{\mathrm{KF}}(I_{\mathrm{GT}}, I_{\mathrm{base}}), \qquad \mathbf{g}^{G}_{\mathrm{text}} = \nabla_{\mathrm{text}}(G, \mathcal{F}^{\mathrm{KF}}_{G}), \tag{19}$$

where I_{base} is the current reconstructed keyframe and every $\xi \in \mathcal{F}^{\mathrm{KF}}_{G}$ records the dimension, image location, GT observation, and current-reconstruction observation of one grounded difference. Thus, the modification direction comes from direct differences, while Eq. 18 supplies an independent verification score.

Inner loop, shot setting. The fixed shot-reward checklist supplies both the failed items $\mathcal{F}^{\mathrm{shot}}_{G}$ and the score $R^{\mathrm{shot}}_{\mathrm{reward}}$, defined in Sec. D.2.

Outer loop. We describe the P_{shot} branch; the P_{kf} branch is identical with the keyframes of the selected shots and their image QA banks in place of shots and video QA banks. Let $\mathcal{S}_i$ be the set of selected shots of pseudo-training video V_i (Algorithm 2), and let $\mathcal{Q}_i = \{\mathcal{Q}(x_{i,s}) : s \in \mathcal{S}_i\}$ collect their frozen banks. Under policy P, video V_i is encoded by the Agentic Video Encoder (Eq. 4), and $\hat{x}_{i,s}(P)$ denotes the reconstruction of selected shot s. $\mathrm{Assess}(P, \mathcal{S}_i, \mathcal{Q}_i)$ returns

$$\begin{aligned} R_i(P) &:= \frac{1}{|\mathcal{S}_i|} \sum_{s \in \mathcal{S}_i} R^{\mathrm{QA}}_{\mathrm{reward}}(x_{i,s}, \hat{x}_{i,s}(P)), \\ \mathcal{F}_P &:= \bigcup_{s \in \mathcal{S}_i} \mathcal{F}_{\mathrm{QA}}(x_{i,s}, \hat{x}_{i,s}(P)), \end{aligned} \tag{20}$$

so the video-level reward is the arithmetic mean of shot-level pass rates, not a pooled pass rate over all questions. This is an implementation-level expansion of the quantities in Algorithm 2: $R_c = R_i(P)$, $R'_c = R_i(P')$, and $\mathcal{F}_{\mathrm{QA}}$ in the algorithm is $\mathcal{F}_P$. The historical rewards R'_h and R^{ref}_h apply the same equal-weight average to replayed samples under P' and P_{ref} (Sec. D.3). Throughout pseudo-training, the question banks, and hence their dimension composition, remain fixed. In every setting, R_{reward} drives optimization, while R_{eval} stays outside both loops and is used only for final reporting.

D Dual-Loop Acceptance Gates

This section instantiates the Anti-Degradation Gate of the inner loop (Sec. 4.3.3, Algorithm 1) and the Anti-Forgetting Gate of the outer loop (Sec. 4.3.2, Algorithm 2), and then specifies the sequential

pseudo-training procedure. Every gate is a conjunction of an improvement check and a preservation check:

$$\Gamma^{\beta}_{\text{inner}} = \text{Improve}_{\beta} \wedge \text{Preserve}_{\beta}, \quad \beta \in \{\text{KF}, \text{shot}\}, \qquad \Gamma_{\text{outer}} = \Gamma^{\text{current}}_{\text{outer}} \wedge \Gamma^{\text{replay}}_{\text{outer}}. \tag{21}$$

Subscripts base and cand denote the current solution and the challenger. All thresholds below are those used in the reported experiments and are absolute differences on the $[0, 1]$ reward scale.

D.1 Keyframe KG Representation Refinement

Let I_{GT}, I_{base}, and I_{cand} denote the GT, current, and challenger keyframes, and let $\Delta R^{\text{KF}}_{\text{reward}} := R^{\text{KF}}_{\text{reward}}(I_{\text{cand}}) - R^{\text{KF}}_{\text{reward}}(I_{\text{base}})$ be the change in the composite keyframe fidelity of Eq. 18; in Algorithm 1, $u = I_{\text{base}}$ and $u' = I_{\text{cand}}$. A pairwise evaluator compares the two reconstructions against I_{GT} twice, reversing their presentation order. Let $\text{Closer}(I_{\text{GT}}; I_1, I_2)$ return I_1, I_2, or a tie. We abbreviate order-consistent agreement by

$$\text{PairCons}(I_{\text{GT}}, I_{\text{base}}, I_{\text{cand}}) := \mathbb{I}\begin{bmatrix} \text{Closer}(I_{\text{GT}}; I_{\text{base}}, I_{\text{cand}}) = I_{\text{cand}} \\ \wedge\, \text{Closer}(I_{\text{GT}}; I_{\text{cand}}, I_{\text{base}}) = I_{\text{cand}} \end{bmatrix}. \tag{22}$$

$\text{PairCons} = 1$ means that the challenger is judged closer to GT under both presentation orders; a disagreement or tie returns zero. The two checks of the keyframe gate are

$$\text{Improve}_{\text{KF}} := \text{PairCons}(I_{\text{GT}}, I_{\text{base}}, I_{\text{cand}}), \qquad \text{Preserve}_{\text{KF}} := \mathbb{I}\big[\Delta R^{\text{KF}}_{\text{reward}} \geq -\epsilon_{\text{KF}}\big], \tag{23}$$

with $\epsilon_{\text{KF}} = 0.05$. $\text{Improve}_{\text{KF}}$ requires an order-consistent visual improvement, and $\text{Preserve}_{\text{KF}}$ limits regressions in composite fidelity to ϵ_{KF}. The bounded tolerance leaves self-evolution room to accept revisions that improve the keyframe overall at a small cost in individual aspects. We use $T_{\text{inner}} = 4$ rounds per keyframe and stop early after two consecutive rejected rounds. Because each accepted round is compared only with the preceding solution, a chain of pairwise wins can drift; before returning, the refined keyframe is therefore compared once more with the initial reconstruction by the same double-order evaluator, and the refinement is reverted if the initial reconstruction is judged closer to GT. The complete pairwise judgment prompt is reproduced in Sec. N.

D.2 Video-Shot KG Representation Refinement

For $x \in \{\text{base}, \text{cand}\}$, let V be the GT shot and $\hat{V}_x$ its reconstruction; in Algorithm 1, $u = \hat{V}_{\text{base}}$ and $u' = \hat{V}_{\text{cand}}$.

Shot reward. For each of the $D_{\text{opt}} = 8$ reward dimensions, the shot-reward checklist lists atomic facts from the GT shot and flags those critical to the dimension. The frozen judge then labels each fact in $\hat{V}_x$ as match, partial, mismatch, or absent, citing what it observes. The judge never outputs a score. Instead, the dimension score $\rho^{\text{shot}}_{x,d} \in [0, 1]$ is computed by rule as the mean credit over judged items (match $= 1$, partial $= 0.5$, mismatch $= 0$), with absent items excluded from the count. A single mismatched critical item caps the score at 0.4, so a severe error cannot be averaged away by many minor matches. The shot reward is the uniform dimension mean

$$R^{\text{shot}}_{\text{reward},x} := R^{\text{shot}}_{\text{reward}}(V, \hat{V}_x) := \frac{1}{D_{\text{opt}}} \sum_{d=1}^{D_{\text{opt}}} \rho^{\text{shot}}_{x,d}, \tag{24}$$

and the mismatched or absent items form the failure set $\mathcal{F}^{\text{shot}}_{G}$.

Target selection. Each round repairs one dimension. The target $d^{\star}$ is the lowest-scoring dimension of the current reconstruction among those scoring below 0.75, excluding any dimension on which two challengers have already been rejected. The failed items of $d^{\star}$, critical ones first, serve as the evidence for the textual gradient.

Acceptance gate. A challenger is accepted when it passes both checks:

$$\begin{aligned}\mathrm{Improve}_{\mathrm{shot}} &:= \mathbb{I}\big[R^{\mathrm{shot}}_{\mathrm{reward,cand}} > R^{\mathrm{shot}}_{\mathrm{reward,base}} + 0.02\big]\,,\\ \mathrm{Preserve}_{\mathrm{shot}} &:= \mathbb{I}\left[\sum_{d\neq d^\star} \max\big(0, \rho^{\mathrm{shot}}_{\mathrm{base},d} - \rho^{\mathrm{shot}}_{\mathrm{cand},d}\big) \le 0.15\right].\end{aligned} \tag{25}$$

$\mathrm{Improve}_{\mathrm{shot}}$ is measured on the overall shot reward rather than on $d^\star$ alone, because a revision aimed at $d^\star$ often recovers facts in other dimensions as well. $\mathrm{Preserve}_{\mathrm{shot}}$ bounds the total regression across the non-target dimensions, leaving room for revisions that trade small local losses for a clear overall gain. We use $T_{\mathrm{inner}} = 3$ rounds per shot and stop early after two consecutive rejected rounds.

D.3 Encoding Policy Acceptance Gate

The outer-loop gate of Algorithm 2 uses $\delta = 0.02$ and $\epsilon_{\mathrm{hist}} = 0.05$, with R_c and R_h computed as in Eq. 20. Both checks are specified below.

Current-video check. The challenger must improve the current-video reward by more than δ:

$$\Gamma^{\mathrm{current}}_{\mathrm{outer}} = \mathbb{I}[R'_c > R_c + \delta]\,. \tag{26}$$

Historical replay check. After each pseudo-training video, its samples and frozen checks are added to the replay memory $\mathcal{M}$ (Algorithm 2). $\mathrm{Replay}(\mathcal{M}, m, n)$ covers every preceding video in $\mathcal{M}$: from each preceding video, the P_{shot} branch uniformly samples $m = 3$ shots with their frozen video-QA banks, and the P_{kf} branch uniformly samples $n = 5$ keyframes with their frozen image-QA banks. Both the challenger P' and the reference P_{ref}, frozen at the start of the current video, are re-run on the same sampled records; averaging their rewards with equal weight as in Eq. 20 gives R'_h and R^{ref}_h, and

$$\Gamma^{\mathrm{replay}}_{\mathrm{outer}} = \mathbb{I}\big[R'_h \ge R^{\mathrm{ref}}_h - \epsilon_{\mathrm{hist}}\big]\,. \tag{27}$$

Comparing against P_{ref} rather than the current policy bounds cumulative regression across all updates on the current video. For the first pseudo-training video, $\mathcal{M}$ is empty and $\Gamma^{\mathrm{replay}}_{\mathrm{outer}} = 1$.

D.4 Encoding Policy Pseudo-Training Procedure

Pseudo-training runs two separate prompt-rewriting branches, one for P_{shot} and one for P_{kf}, over $L = 6$ videos from $\mathcal{D} = \{V_i\}_{i=1}^{L}$. Following the policy-level objective in main-paper Eq. 3, videos are drawn uniformly from $\mathcal{D}$ and then processed one at a time. In each branch, a video is processed as follows: (i) the previously accepted prompt is inherited and frozen as P_{ref}; (ii) three shots are selected ($|\mathcal{S}_i| = 3$) and their QA banks are frozen; (iii) the current policy is re-evaluated to obtain R_c; and (iv) $T_{\mathrm{outer}} = 3$ challengers are proposed and gated as in Algorithm 2, where each accepted challenger becomes the current policy for the next round. The final prompt is passed to the next video, and the video's samples are added to $\mathcal{M}$. After the sixth video, the two accepted prompts are assembled with the fixed P_{film} into the pseudo-trained policy P^*. No foundation-model weights are updated.

E Agentic Video Encoder Implementation Details

This section details the Agentic Video Encoder of main-paper Sec. 4.1: preprocessing, followed by its two encoding steps, hierarchical agentic understanding (Eq. 4) and deterministic Film KG assembly (Eq. 5). Preprocessing is fixed and separate from the encoding policy: neither P_{shot} nor P_{kf} performs temporal segmentation.

E.1 Shot Segmentation and Keyframe Selection

Preprocessing consists of two fixed operators: Seg partitions V into shots, and Key selects keyframes within each shot. For Seg, a fixed video-understanding model inspects the complete source video and proposes ordered shot boundaries using three forms of evidence: changes in visual semantics, explicit cinematic-transition cues, and the completion boundaries of spoken utterances. The segmentation

prompt targets shots of approximately 10 seconds and normally no more than 15 seconds, but an utterance is kept intact even when doing so requires a longer shot. A deterministic parser converts the returned timestamps into contiguous frame-index intervals $\{\mathcal{I}_i\}_{i=1}^{S}$, where $\mathcal{I}_i$ is the frame interval of shot s_i, after which FFmpeg performs frame-accurate slicing. Thus, the macro segmentation is semantic and audiovisual; the frame-difference and motion-stability calculations described next belong to keyframe selection within an already segmented shot.

Key is an adaptive, model-free computer-vision operator that returns the keyframe set $\mathcal{K}_i := \{f_{i,j}^{\mathrm{kf}}\}_{j=1}^{J_i}$ of shot s_i. Within each s_i, it computes the mean grayscale difference between every pair of adjacent frames. Its change threshold is adapted to that shot as the larger of a fixed floor of 40 and the median difference plus eight robust standard deviations, where the robust deviation is 1.4826 times the median absolute deviation. Non-maximum suppression retains only the strongest change within a 0.5-second neighborhood; intervals shorter than 0.8 seconds and neighboring boundaries separated by less than 0.4 seconds are merged. For every remaining interval, the selector discards the outer 15% transition margins, retains the calmest 40% of candidate frames by local motion energy, and chooses the sharpest of those frames by Laplacian variance. If too few candidates remain, it falls back to the interval midpoint. The resulting number J_i is content-adaptive and is uniformly downsampled when necessary to enforce $J_{\max} = 5$, matching the shared per-shot decoder-interface budget of at most five generated reference keyframes (Sec. G). All steps in Key use only video pixels and deterministic computer-vision tools; they make no model call.

E.2 Three-Level Specialized Understanding and Inter-Level Context Injection

The three levels below instantiate $\mathrm{Understand}(\cdot; P_b)$ in main-paper Eq. 4. Film-level understanding, $\mathrm{Understand}(V; P_{\mathrm{film}})$, processes the full video and returns the film-level textual result $\mathcal{C}_{\mathrm{film}}$ and the shared registry $\mathcal{B}_{\mathrm{reg}} = \mathcal{B}_{\mathrm{char}} \cup \mathcal{B}_{\mathrm{scene}} \cup \mathcal{B}_{\mathrm{obj}}$. The former records global narrative, style, ordered-shot relations, and other whole-film evidence. The latter assigns stable names, IDs, appearance descriptions, and variants to recurring characters, scenes, and objects. These IDs form the common anchors used by all later textual records and by graph-addressable asset indices.

Shot-level understanding, $\mathrm{Understand}(s_i, \mathcal{C}_{\mathrm{film}}, \mathcal{B}_{\mathrm{reg}}; P_{\mathrm{shot}})$, receives the complete shot together with the film-level result and registry. It specializes in shot understanding and writes $\mathcal{C}_{\mathrm{shot},i}$ as an information-description textual representation of the shot's subjects and entity states, spatial and temporal dynamics, camera behavior, style, dialogue, music, and sound. Injecting $\mathcal{C}_{\mathrm{film}}$ supplies global narrative and ordered cross-shot evidence that is not recoverable from s_i alone, while the registry forces recurring entities to reuse canonical names and IDs.

Keyframe-level understanding, $\mathrm{Understand}(f_{i,j}^{\mathrm{kf}}, \mathcal{C}_{\mathrm{shot},i}, \mathcal{B}_{\mathrm{reg}}; P_{\mathrm{kf}})$, receives each keyframe in $\mathcal{K}_i$, the shot-level result $\mathcal{C}_{\mathrm{shot},i}$, and the complete registry. It specializes in keyframe understanding and writes a separate $\mathcal{C}_{\mathrm{kf},i,j}$ for every $f_{i,j}^{\mathrm{kf}} \in \mathcal{K}_i$, covering subjects, static action state, appearance, spatial layout, scene, composition, lighting, color, and style. The complete registry supplies the same canonical entity names and IDs available at the shot stage. The implementation injects only $\mathcal{C}_{\mathrm{shot},i}$ into keyframes belonging to s_i; it never injects another shot's local result. Selected temporal fields from $\mathcal{C}_{\mathrm{shot},i}$ locate a static pose within the full action process and disambiguate evidence that a single frame cannot resolve.

During policy pseudo-training, only P_{shot} and P_{kf} receive textual updates; P_{film}, the preprocessing operators Seg and Key, and all foundation-model weights remain fixed.

After the three understanding stages, $\mathrm{BuildGraph}$ parses their structured textual fields and registry IDs, instantiates the fixed Film KG node schema, and adds hierarchy, temporal-order, state-binding, semantic, and asset-reference edges according to fixed rules. This assembly is deterministic and invokes no additional foundation model. Section K gives the full stored schema and propagation rules.

F Agentic Video Encoder System Prompts

The Ablation Study (Sec. 5.3) compares three complete encoding policies $P = (P_{\mathrm{film}}, P_{\mathrm{shot}}, P_{\mathrm{kf}})$. All three share the same fixed film-level prompt P_{film} and differ in their shot- and keyframe-level components:

- **Initial policy** P_0: the initial shot- and keyframe-level prompts from which pseudo-training starts.
- **Pseudo-trained policy** P^*: the shot- and keyframe-level prompts accepted after the last pseudo-training video (Sec. D.4).
- **Human-tuned policy**: independently and manually engineered shot- and keyframe-level prompts; it is not P_0 and is not produced by pseudo-training. It was manually tuned with the same Qwen-3.7-Plus understanding model used by all policies.

Section N provides complete direct English translations of the Chinese prompts: the fixed film-level prompts for whole-film understanding and for character, scene, and object registry construction; the initial and pseudo-trained shot- and keyframe-level prompts; and the human-tuned shot- and keyframe-level prompts.

G Baseline Adaptation and Fairness

This section specifies how the baselines in the Experimental Setup (Sec. 5.1) and Reconstruction Fidelity (Sec. 5.2) are connected to the shared reconstruction pipeline.

We compare with VideoAnalyzer (Docusphere, 2026), Storyboard Studio (BroderQi, 2026), and soap2soap (Song et al., 2026). Because these systems do not all include an end-to-end film-reconstruction decoder, we map each method's own representation to the same fixed image and video generators, preserving each method's analysis and representation.

VideoAnalyzer. We retain its temporal scene analysis, aligned transcript, extracted frames, and video representation. Its own segments and descriptions are converted into reconstruction prompts for the shared generators.

Storyboard Studio. We retain its original shot analysis, contact sheets, and storyboard structure.

soap2soap. We retain its video analysis, character representation, prompt construction, and keyframe stages. These outputs are passed to the shared generators, and its reconstruction uses only assets produced by soap2soap.

Source-pixel exclusion. All four systems obey the same strict boundary between representation construction and reconstruction generation. A system may inspect the source video when constructing its representation, and source frames may be used as GT only during evaluation; however, no source-video frame, crop, screenshot, or other source-pixel image is passed to either generation model or used as a visual reference for reconstruction. Every reconstructed keyframe is newly generated by the shared image generator from the text-centered structured description produced by the corresponding representation system. Any such keyframe subsequently used by the video generator remains a generated reconstruction asset rather than copied source content. Accordingly, the extracted frames retained by VideoAnalyzer and the contact sheets retained by Storyboard Studio are analysis inputs only and never generation inputs. Reconstruction quality therefore measures the information carried by each system's text-centered representation, the form that agents can directly read, reason over, and edit.

Shared per-shot decoder-interface budget. All four systems use the same shared per-shot decoder-interface budget from the Experimental Setup (Sec. 5.1): at most 5 newly generated reference keyframes and 1,200 shot-prompt tokens per shot. They constrain only what the decoder receives, not the source material available during representation construction and evaluation.

Shared models. All four systems receive the same source videos and use the same Qwen-3.7-Plus understanding model (Alibaba Cloud, 2026b), fixed Nano Banana Pro image generator (Google AI, 2026b), HappyHorse 1.0 reference-to-video generator (Alibaba Cloud, 2026a), and evaluator, so the comparison does not depend on different understanding or generation backbones.

In a pilot with a second video generator, Seedance 2.0 (ByteDance Seed Team, 2026), a subset of shots was regenerated with the representations, prompts, and evaluator held fixed; the Overall reconstruction score changed by at most 1.22 percentage points.

H Dataset and Reproducibility Details

This section describes the released Film KG dataset and the six pseudo-training and 18 evaluation videos used in the Experimental Setup (Sec. 5.1).

H.1 Released Film KG Dataset

We construct and release a dataset built from about 500 videos of 2–20 minutes each, comprising more than 20K shots and more than 50K keyframes. It contains the text portion of Film KG representations derived from high-quality, human-made film content, including fine-grained structured descriptions of scripts, characters, scenes, objects, shots, and keyframes. These representations retain enough information to support reconstruction of the source film content through a fixed generation pipeline. The input videos are collected from open-source online data and include high-quality film content, award-winning short films, and popular human-produced AI videos.

The release layer of each Film KG $G = (\mathcal{N}_G, \mathcal{E}_G, \mathcal{A}_G)$ consists of its structured-text nodes $\mathcal{N}_G$ and relations $\mathcal{E}_G$; the generated assets $\mathcal{A}_G$ are not distributed but can be re-rendered from the text nodes. Users can connect their own image-, audio-, and video-generation APIs, or feed the representations directly into AVA-Encoder, to render the corresponding assets. The graph structure also supports linked editing in which a node update changes its dependent video content. Without rendering any video, users can further use the structured text nodes and their dependencies as high-quality reverse-engineered creation records.

H.2 Pseudo-Training and Evaluation Video Collections

Both collections are drawn from open-source data.

Pseudo-training videos. The six-video pseudo-training collection contains the following clips:

1. *The Big Bang Theory*: Sheldon learns Chinese.
2. *The Truman Show*: ending sequence.
3. *Friends*: Rachel's runaway-wedding sequence.
4. Flash's document-stamping sequence.
5. Platform Nine and Three-Quarters sequence.
6. *The Pursuit of Happyness*: interview sequence.

Evaluation videos. The 18-video evaluation collection contains the following clips:

1. *Spirited Away*: Chihiro's parents turn into pigs.
2. *Zombie Cleaner*, an AI-generated short film.
3. Diagon Alley sequence.
4. *Joker*: subway sequence.
5. *Kung Fu Panda*: Master Oogway's classic speech.
6. *Genshin Impact*: Natlan "Little Friend" sequence.
7. *Dream of the Red Chamber*: Baoyu and Daiyu's first meeting.
8. *Rather Than Suspicion*.
9. *Shameless*: selected sequence.
10. *Titanic*: deck-embrace sequence.
11. *Zootopia 2*: selected sequence.
12. Paw-shaped popsicle sequence.
13. *The Secret Life of Walter Mitty*: mountain sequence.
14. *The Million Pound Note*: selected sequence.
15. *3 Idiots*: classroom sequence.
16. *Green Book*: car sequence.

Table 7: V direction by dimension (%). Higher is better.

Method	Char.	Scene	Pos.	Motion	Audio	Style	Camera	Narr.	V mean
VideoAnalyzer	32.9	33.5	23.4	19.8	16.9	17.4	42.7	22.4	26.1
Storyboard Studio	15.8	25.1	14.0	9.6	13.4	12.3	30.5	10.1	16.4
soap2soap	40.9	45.1	34.5	30.2	31.6	28.2	47.7	35.5	36.7
AVA-Encoder	**56.2**	**80.9**	**63.4**	**43.7**	**52.5**	**39.5**	**71.6**	**54.9**	**57.8**

Table 8: KF direction by its seven applicable dimensions (%). Audio is N/A; higher is better.

Method	Char.	Scene	Pos.	Motion	Style	Camera	Narr.	KF mean
VideoAnalyzer	30.0	24.1	24.2	25.9	15.3	60.0	19.9	28.5
Storyboard Studio	27.1	22.5	21.4	30.4	10.1	66.0	22.4	28.6
soap2soap	38.4	38.8	31.6	35.6	27.2	66.2	38.7	39.5
AVA-Encoder	**63.3**	**80.3**	**84.5**	**68.1**	**45.2**	**91.7**	**83.0**	**73.7**

17. *Forrest Gump*: chase sequence.
18. Flying-lesson sequence in which Malfoy causes trouble.

The evaluation collection spans varied content and forms, with 129 annotated shots and 246 selected keyframes.

The six pseudo-training videos and the 18 evaluation videos are disjoint; the pseudo-training procedure is specified in Sec. D.4.

I Fine-Grained Reconstruction Results

Tables 7 and 8 break down the direct V and KF scores of Table 1 (Reconstruction Fidelity, Sec. 5.2) by dimension, computed with Eq. 14. The row means equal the V and KF entries in Table 1.

J Ablation Definitions and Effect Sizes

This section defines every configuration in Table 2 (Ablation Study, Sec. 5.3) and reports the main effect sizes.

Single-level encoder. Single-level understanding replaces the hierarchical film–shot–keyframe encoder; neither optimization loop is used.

w/o both loops. The initial policy P_0 is used with neither policy pseudo-training nor KG refinement.

Human-tuned policy. An independently and manually engineered complete policy (Sec. F), distinct from P_0, is used with neither policy pseudo-training nor KG refinement.

w/o policy pseudo-training. The initial policy P_0 is fixed; KG refinement and its Anti-Degradation Gate remain active.

w/o KG refinement. The pseudo-trained policy P^* is used, and input-specific keyframe and shot refinement is disabled.

w/o inner-loop gate. The pseudo-trained policy P^* is used, and KG refinement runs without the Anti-Degradation Gate, so every KG-refinement challenger replaces the current Film KG.

w/o outer-loop gate. Pseudo-training is re-run from the initial policy P_0 with the Anti-Forgetting Gate disabled, so every policy challenger replaces the current policy. This yields an ungated policy $\tilde{P}$, which is then used with KG refinement and its Anti-Degradation Gate.

w/o both gates. The ungated policy $\tilde{P}$ from the previous variant is used, and KG refinement also runs without the Anti-Degradation Gate.

AVA-Encoder (full). Hierarchical encoding with the pseudo-trained policy P^*, KG refinement, and both gates.

With Data-Dependent KG Representation Refinement disabled in both conditions, the pseudo-trained Agentic Video Encoder policy improves over its independently human-tuned counterpart from 44.4 to 45.8: an **absolute gain of 1.4 percentage points** and a $1.4/44.4 = \mathbf{3.2}\%$ relative improvement. The pseudo-trained policy is also more compact than the human-tuned policy: its shot-level system prompt has 8,052 tokens instead of 31,336 (74.3% fewer), and its keyframe-level system prompt has 4,062 tokens instead of 13,574 (70.1% fewer). Enabling optional test-time Data-Dependent KG Representation Refinement yields the complete AVA-Encoder score of 49.0. The complete configuration improves over the no-loop configuration from 42.4 to 49.0: an **absolute gain of 6.6 percentage points** and a $6.6/42.4 = \mathbf{15.6}\%$ relative improvement. Against the strongest external baseline, soap2soap at 28.3, the 49.0 score is an absolute gain of **20.7 percentage points** and a relative improvement of $20.7/28.3 = \mathbf{73.1}\%$.

K Film KG Representation and Editing

This section specifies the Film KG $G = (\mathcal{N}_G, \mathcal{E}_G, \mathcal{A}_G)$ of main-paper Sec. 4.2: the stored schema, and how the encoder's records are assembled into it.

K.1 Stored Schema

Node types. The ten textual node types of main-paper Sec. 4.2 are serialized as follows: Story, Event, Shot, and Keyframe as `story`, `event`, `shot`, and `keyframe`; Character, Scene, Object, Style, Camera, and Audio as the shot-specific state types `char_state`, `scene_state`, `object_state`, `style_state`, `camera_state`, and `audio_state`. Every node stores a structured text description. A `story` node describes the global narrative, an `event` node describes a local narrative unit, and a `shot` node describes one film shot. The six state types describe their shot-specific character, scene, object, style, camera, and audio information. In particular, `audio_state` describes spoken content and speaker identity, voice properties, background music, sound effects, and audiovisual synchronization in text rather than storing audio data.

Each `keyframe` node additionally stores its image-generation payload. For a node type τ, let $\mathcal{N}_G^\tau \subseteq \mathcal{N}_G$ denote the node instances of that type, so that $\mathcal{N}_G$ is the disjoint union of the ten instance subsets.

Assets. The asset layer $\mathcal{A}_G$ stores or references generated character, scene, and object reference images; generated keyframe images; audio or voice assets; and rendered shot videos. Each asset is connected to the node it was generated from by a typed asset edge in $\mathcal{E}_G$; for example, each keyframe node has an asset edge to its generated image. Edits propagate over nodes, and the assets reached through asset edges from affected nodes are regenerated.

The source-pixel exclusion defined in Sec. G makes the graph an explicit text bottleneck: AVA-Encoder cannot obtain reconstruction fidelity by copying pixels from the source video, but must express reconstruction-critical information through structured text nodes and their relations. Assets in $\mathcal{A}_G$ are generated from those descriptions rather than copied from the source video. The resulting text-centered graph can therefore be directly understood and learned from by downstream agents, queried by entity, event, shot, or asset, edited at a selected node, and reused to generate new multimodal assets.

Relation types. Main-paper Sec. 4.2 groups typed edges into structural, asset, temporal, and semantic dependencies. Asset dependencies are the node–asset edges described above; among nodes, we specify the 11 serialized relation types below. `contains` represents the story–event–shot hierarchy; `binds` links a shot to the states and keyframes used by that shot; and `transition` links successive states of the same entity. `sequence` records the temporal order of shots or events, while `jump` links non-adjacent reappearances of the same character, scene, or object. `spoken_by`

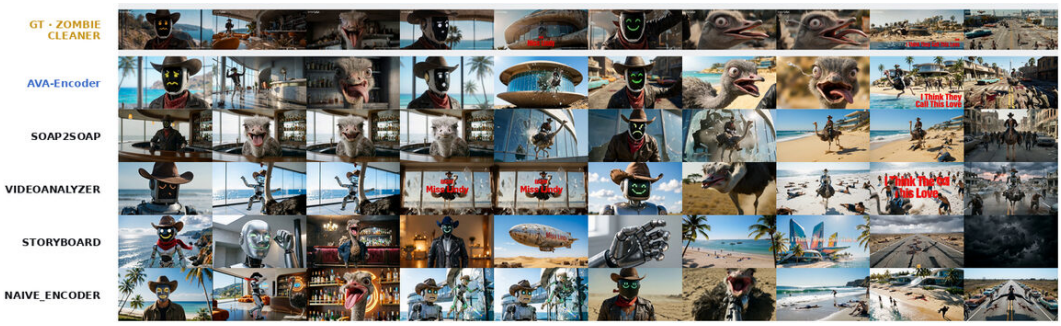


Figure 5: Qualitative reconstruction comparison on *Zombie Cleaner*. Rows show GT, AVA-Encoder, soap2soap, VideoAnalyzer, Storyboard Studio, and the single-level encoder.

links spoken content in an audio state to a character; `rel` records character relationships; `similar` marks visually confusable characters, scenes, or objects; and `features` links a scene to its recurring characters or objects. `narrative` records causal, setup, and callback relations between shots, and `references` records which character or scene occurrences were used to produce a keyframe. Style reaches keyframes through `binds`: a shot binds both its `style_state` and its keyframes. Consistency among states within the same shot is not stored as an edge; it is derived at edit time from shot co-membership.

K.2 Construction and Propagation

Film-level understanding produces three explicit registry banks: a character registry $\mathcal{B}_{\text{char}}$, a scene registry $\mathcal{B}_{\text{scene}}$, and an object registry $\mathcal{B}_{\text{obj}}$. We denote their union by

$$\mathcal{B}_{\text{reg}} := \mathcal{B}_{\text{char}} \cup \mathcal{B}_{\text{scene}} \cup \mathcal{B}_{\text{obj}}. \tag{28}$$

The character registry stores textual descriptions of stable identities and appearance variants, the scene registry stores textual descriptions of recurring environments and their spatial and visual properties, and the object registry stores textual descriptions of recurring props and their observable states. Any generated reference images associated with these entries remain in $\mathcal{A}_G$. The registry is injected into shot- and keyframe-level understanding as described in Sec. E.2, binding observations at both levels to stable identities across cuts. Each object-registry entry is instantiated as an `object_state` in every shot where that object appears. A `binds` edge attaches the state to its shot, through which it reaches the shot's keyframes, and `transition` links consecutive states of the same object. Film-level understanding, the three registry banks, shot-level understanding, and all keyframe-level understanding results are then combined into structured records. The nodes and relations above are constructed deterministically from those records; graph construction requires no additional model call. How edits propagate through the resulting graph is specified in Sec. L.2.

L Additional Qualitative Results

This section provides qualitative reconstruction comparisons for Reconstruction Fidelity (Sec. 5.2) and graph-based editing examples for Representation Applications (Sec. 5.4).

L.1 Reconstruction Comparisons

Figures 5–10 compare film reconstruction across six source videos. In each comparison, rows show GT, AVA-Encoder, soap2soap, VideoAnalyzer, Storyboard Studio, and the single-level encoder (the ablation in Table 2) under the shared generation setting.

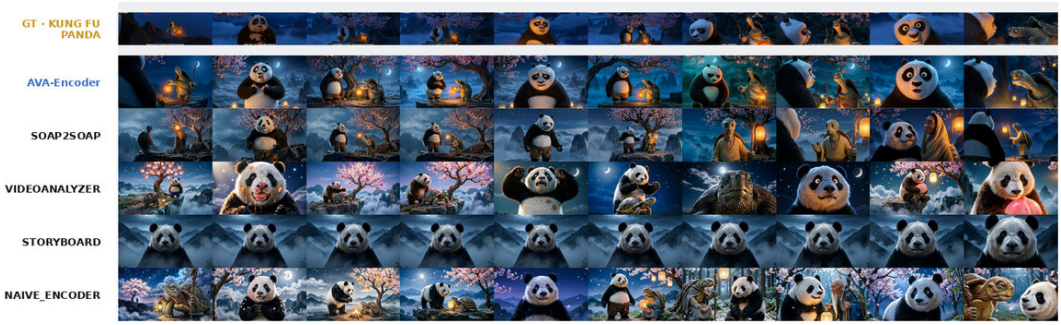


Figure 6: Qualitative reconstruction comparison on *Kung Fu Panda*, using the same row order as in Fig. 5.



Figure 7: Qualitative reconstruction comparison on *Shameless*, using the same row order as in Fig. 5.



Figure 8: Qualitative reconstruction comparison on *Titanic*, using the same row order as in Fig. 5.

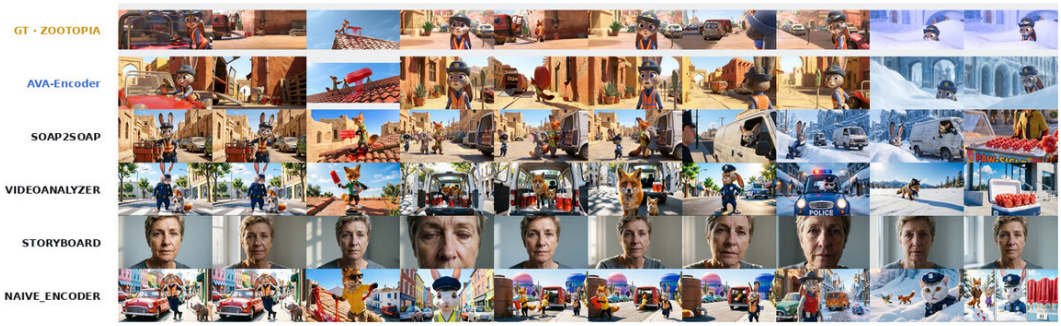


Figure 9: Qualitative reconstruction comparison on the paw-shaped popsicle sequence, using the same row order as in Fig. 5.



Figure 10: Qualitative reconstruction comparison on the flying-lesson sequence in which Malfoy causes trouble, using the same row order as in Fig. 5.

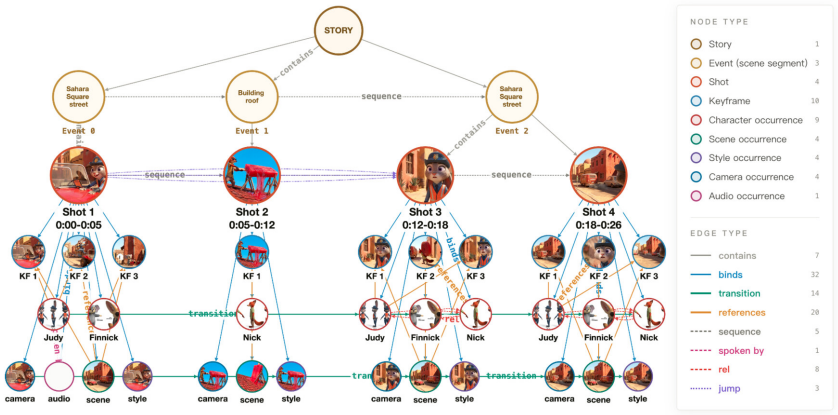


Figure 11: Film KG visualization. The story–event–shot hierarchy appears above the shot-bound state and keyframe nodes. Typed edges expose temporal, semantic, and asset-reference dependencies that are not represented by a flat shot list. The visualization shows the node and relation types instantiated in this clip; object states and the remaining relation types of Sec. K do not occur here.

L.2 Graph-Topology Editing

This subsection implements the edit scope of main-paper Eq. 6 used for graph-based editing (Sec. 5.4). Given edited nodes $Z \subseteq \mathcal{N}_G$, it determines the affected subgraph identified by the edit-conditioned propagation rules; assets in $\mathcal{A}_G$ reached through asset edges from the affected nodes are then regenerated. An edit is first assigned one of four edit types e: identity replacement, content rewrite, visual-treatment change, or parameter adjustment. The edit type determines which typed edges may carry the edit and in which direction.

Figure 11 visualizes the stored graph described in Sec. K. Figures 12 and 13 show how the same graph is inspected during editing. The overview retains the complete film hierarchy and all typed layers; selecting one asset isolates the dependency paths that determine which linked states and rendered outputs may require an update.

The main-paper identity example (Figure 4) and the additional classical-drama identity and visual-treatment examples below (Figures 14 and 15) illustrate the two principal edit cases. The rules below formalize how their linked states and rendered assets are updated; at the graph level, nodes outside the affected subgraph are left unchanged. In the rendered outputs of the shown examples, unrelated characters, scenes, and compositions are also preserved.

Table 9 summarizes the propagation rule and resulting update for each edit type.

Table 9: Edit-type-specific propagation used for graph-topology editing.

Edit type	Propagation and resulting update
Identity	Follow entity-state transitions and asset references; refresh linked identities, keyframes, and shots.
Content	Follow the edited shot's bindings; recompile its prompts, rerender its assets, and check one-step semantic dependents.
Visual treatment	Follow style-state transitions and shot bindings; apply the treatment consistently to linked keyframes and shots.
Parameter	Follow the edited state and its shot binding; update only the selected camera, lighting, scene, or object attributes.

The affected subgraph is determined in three stages, corresponding to BFS_e, Check_e^1, and Up in main-paper Eq. 6.

Material closure (BFS_e). The first stage collects nodes whose linked assets must be regenerated together. State transitions connect occurrences of the same entity, `references` edges connect character and scene occurrences to their keyframes, and `binds` edges connect a shot to its states and keyframes, through which a style change reaches the shot's keyframes. These dependencies are followed repeatedly until no additional node is reached.

Semantic consistency check (Check_e^1). The second stage adds one-hop contextual dependencies, drawn from stored semantic edges and from within-shot co-membership derived at edit time, that may require a local revision after the material update, including within-shot character–scene compatibility, audio ownership, camera framing, temporal order, and narrative relations. This one-hop check improves global consistency while keeping each edit local.

Hierarchical update (Up). The third stage adds the event and story nodes that contain an affected shot, so their summaries can be revised without changing unaffected branches of the graph hierarchy. Formally, let $\mathcal{E}_e^{\mathrm{mat}}$ and $\mathcal{E}_e^{\mathrm{sem}}$ be the directed material and semantic dependency sets allowed for edit type e; $\mathcal{E}_e^{\mathrm{sem}}$ includes the derived within-shot relations and is therefore not a subset of the serialized edge set $\mathcal{E}_G$, and let $\mathrm{Dep}_e^{\nu}(Z) := \{v \mid \exists u \in Z,\ (u,v) \in \mathcal{E}_e^{\nu}\}$ return the one-step dependents of a node set Z for $\nu \in \{\mathrm{mat}, \mathrm{sem}\}$. The three operators of main-paper Eq. 6 are implemented as

$$B^{(0)} := Z, \qquad B^{(\ell+1)} := B^{(\ell)} \cup \mathrm{Dep}_e^{\mathrm{mat}}\left(B^{(\ell)}\right), \qquad \mathrm{BFS}_e(G, Z) := \bigcup_{\ell \geq 0} B^{(\ell)},$$
$$\mathrm{Check}_e^1(B) := \mathrm{Dep}_e^{\mathrm{sem}}(B), \tag{29}$$

where ℓ is the propagation depth, and Up adds at most the two enclosing hierarchy levels of each affected shot, its event and story. $\mathcal{N}_{\mathrm{aff}} = \mathrm{Up}(B \cup \mathrm{Check}_e^1(B))$ is then the affected node set identified by these edit-conditioned propagation rules.

Each affected node receives an update action and the dependency path that caused it to be selected. If several paths reach the same node, their reasons are combined and the action that performs the more complete update is retained. Identity references and text states are updated first; generation prompts are then compiled from the updated states, followed by keyframe rendering and video rendering. The original graph is retained, and edits are represented as an overlay, so unrelated nodes and the source representation remain unchanged at the graph level.

Propagation by edit type. For identity replacement, Z contains the selected character state; an edit to a shared registry entry is first mapped to the set of `char_state` nodes instantiated from that entry. The material closure follows state transitions in both temporal directions to reach the character's other occurrences, then follows asset references to the corresponding keyframes and shot bindings to the shots that must be rerendered. The one-step semantic check examines camera framing and audio ownership because a changed body shape or speaker identity can require a local adjustment, but it does not recursively rewrite unrelated audio or camera tracks. Up then records which enclosing event and story summaries require revision.

For a content rewrite, propagation starts from the selected shot node and its bound keyframe nodes, whose prompts and linked assets are regenerated. Temporal-order and narrative relations are checked for one step so that adjacent context can be revised when necessary, without carrying the edit through the full sequence. A visual-treatment edit instead follows the transitions of the same style state and, through each shot's `binds` edges, reaches the corresponding keyframes, propagating the treatment to all linked shots. A parameter edit, such as a camera, lighting, scene, or object adjustment, remains attached to the edited state and its bound assets. The rule therefore improves cross-shot and global consistency for identity and visual-treatment edits while retaining local control for content and parameter changes.

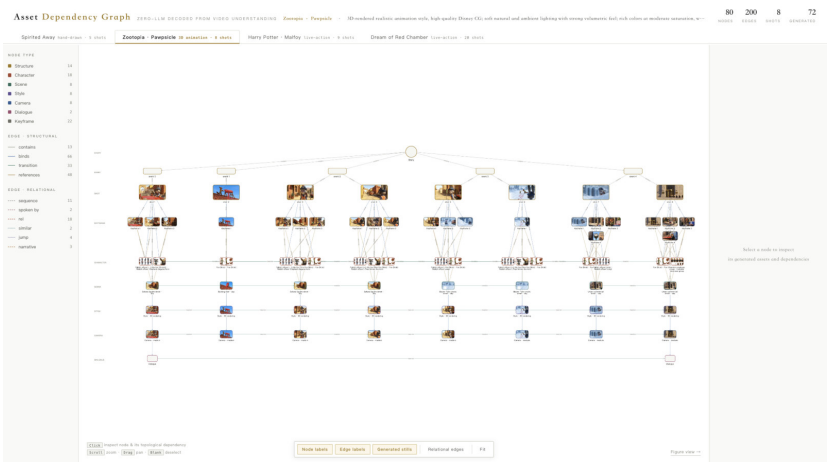


Figure 12: Graph-editing interface before node selection. The complete film graph is organized by hierarchy and semantic layer, allowing an editor to locate a story unit, shot state, or rendered keyframe without flattening its dependencies.

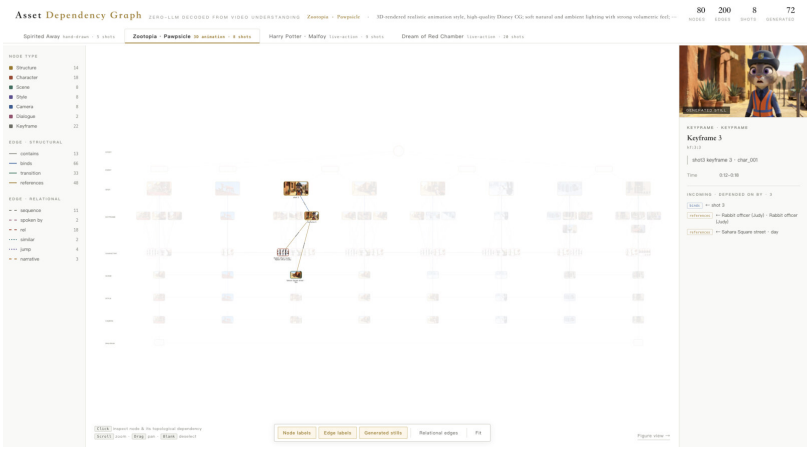


Figure 13: Graph-editing interface after selecting a keyframe node. The interface highlights the selected node's incoming and outgoing dependency paths and displays its stored asset information, making the affected subgraph used by the propagation rule explicit.

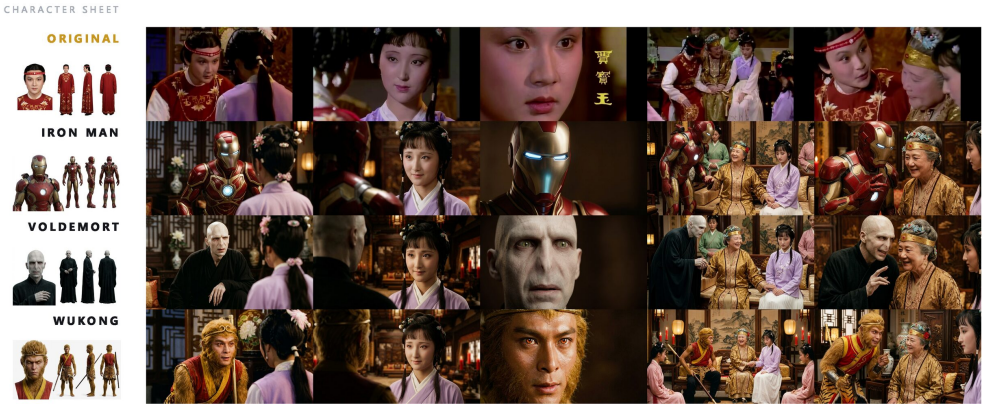


Figure 14: Cross-domain identity replacement in a classical-drama sequence.

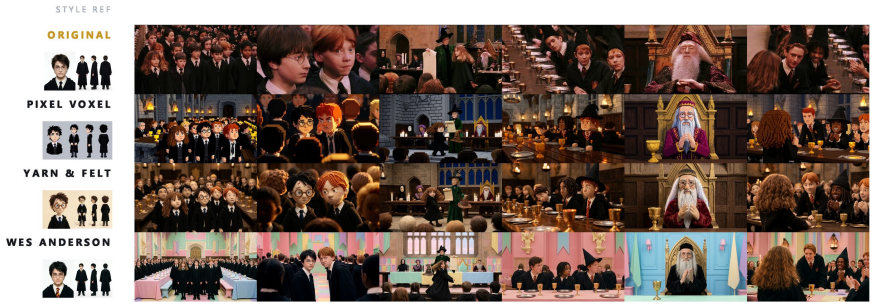


Figure 15: Graph-topology visual-treatment editing. A visual-treatment edit follows the corresponding style-state transitions and shot bindings, applying a consistent treatment to the linked sequence without rewriting each shot independently.

M Downstream Story-Video Evaluation Detail

This section specifies the evaluation behind Table 3 (downstream video generation, Sec. 5.4): the reference-free quality-evaluation rubric, grade-to-score conversion, and the Film KG (text) condition.

The downstream evaluation is a reference-free quality evaluation performed by Gemini-3.1-Pro-Preview: it scores the quality of each generated story video without comparing it to a source film. The five dimensions are Character (appearance, personality, and arc), Plot (structure, causality, tension/pacing, and payoff), Camera (composition, movement, editing, and narrative service), Style (consistency, color/tone, fidelity, and art design), and Audiovisual Quality (completeness, synchronization, music, and mixing). Each sub-dimension receives a grade from A to D, mapped to 4–1; a dimension is the mean of its sub-dimensions and Overall is the unweighted mean of the five dimensions. In the Film KG (text) condition (the "with" score in each cell of Table 3), the complete Film KG is supplied once as text before generation, with only a basic one-sentence request that the framework refer to this representation for the current case. This single, system-independent injection is the only change from the corresponding no-reference condition; the evaluation rubric remains fixed.

N COMPLETE SYSTEM PROMPTS

This section supplements the Agentic Video Encoder (Sec. 4.1), R_{reward} and R_{eval} (Sec. 4.3.1), and Experiment Setup (Sec. 5.1) with complete direct English translations of the Chinese prompts used for reconstruction evaluation, QA-reward construction and answering, keyframe comparison, and encoding. The encoder prompts cover the fixed film-level component (whole-film understanding and registry construction); the initial, pseudo-trained, and human-tuned shot-level components; and the initial, pseudo-trained, and human-tuned keyframe-level components. The experiments use the original Chinese prompts, while these translations document the corresponding experimental specification.

Complete Reconstruction-Evaluation System Prompts

The system prompts used during the reported experiments were written in Chinese; direct English translations are provided here for presentation. All compared methods use the same frozen prompts.

Direct Reconstruction Evaluation

Character

```
You are a strict video-reconstruction consistency judge. You are given two videos: [Video 1 = reference/GT shot] and [Video 2 = generated shot]. They are the "original" and "AI reconstruction" of the same shot. In this evaluation, assess only one dimension: character.

## Protocol (three steps; do not skip any step or inspect Video 2 before making the checklist)
[Step 1: inspect only Video 1] Decompose every main character in the GT into a list of atomic checkpoints. Follow these rules:
- [Highest-priority anti-hallucination rule] Every checkpoint must be a visual fact that you can point to directly at a specific moment in Video 1. Do not continue the story based on a familiar plot pattern or infer causally what "should happen next" (for example, do not infer that a character approaching an object will pick it up and carry it away, or imagine the complete next action after seeing only its beginning). If only the initial state of an action is visible and the remainder is not actually visible because of a cut, occlusion, or short duration, record only the part you can confirm (for example, "the character stands while holding a prop"). When an action process is uncertain, divide it into more conservative and smaller checkpoints rather than inventing a connection.
- Each checkpoint must contain only one directly visible and verifiable fact. Do not combine facts with "and"; convert statements of degree into binary propositions.
- Priority (all P0 items are mandatory) for every main character:
  * Core identity: gender and age group (crit=true)
  * Appearance: at least 1 item (crit=true; face shape, facial features, skin tone, or a salient facial mark, such as "round face, thick eyebrows, and a full beard" or "oval face and straight bangs")
  * Body proportions: at least 1 item (height, build, body shape, or proportions, such as "tall and slim with long legs" or "short, broad, and sturdy")
  * Expression and emotion: at least 1 item (crit=true; be specific: smiling, crying, surprised, calm, shy, etc.)
  * Gaze direction: at least 1 item (where the character looks, whom the character meets eyes with, or whether the head is lowered or raised)
  * Pose and action: at least 1 item (describe the pose and action, with its concrete form in parentheses, such as "half-kneeling in salute (both knees on the ground and upper body leaning forward)" or "sitting upright (hands folded on the lap)")
  * Main clothing form and dominant color: 1-2 items (crit=true; ask only about the principal form and color)
  * Main hairstyle or headwear feature: at most 1 item
- Details such as accessories and patterns must constitute no more than 20% of all items. With multiple characters, identify the character in every item (for example, "the boy in red").
- Scale the total number of items with the number of characters: 8-11 for one character and 14-18 for two characters.

[Step 2: compare each item with Video 2] For every item, provide:
- verdict: "match" / "partial" / "mismatch" / "absent" (the element is missing or not visible in Video 2)
- evidence: one sentence describing the specific observation in Video 2 (for example, "In Video 2, the woman looks straight ahead with a neutral expression"). Do not use empty statements such as "consistent/inconsistent with GT." Verify every item independently; do not mark everything as a match because the overall result appears similar. Real reconstructions usually contain some mismatches, which must be reported faithfully.

[Step 3] Do not output any score. A downstream program computes the score from the verdicts; you are responsible only for facts and evidence.

Return strict JSON only, with no Markdown fence or additional text:
{"dim":"character",
 "checks":[{"sub":"identity|appearance|build|demeanor|gaze|posture|costume|hair","point":"original checkpoint text","crit":true|false,"verdict":"match|partial|mismatch|absent","evidence":"specific observation in Video 2"}, ...],
 "desc_src":"two-sentence summary of the characters in GT","desc_gen":"two-sentence summary of the characters in Video 2"}
```

Scene

```
You are a strict video-reconstruction consistency judge. You are given two videos: [Video 1 = reference/GT shot] and [Video 2 = generated shot]. They are the "original" and "AI reconstruction" of the same shot. In this evaluation, assess only one dimension: scene.

## Protocol (three steps; do not skip any step or inspect Video 2 before making the checklist)
[Step 1: inspect only Video 1] Decompose the GT scene into a list of atomic checkpoints. Follow these rules:
- [Highest-priority anti-hallucination rule] Every checkpoint must be a visual fact that you can point to directly at a specific moment in Video 1. Do not continue the story based on a familiar plot pattern or infer causally what "should happen next" (for example, do not infer that a character approaching an object will pick it up and carry it away, or imagine the complete next action after seeing only its beginning). If only the initial state of an action is visible and the remainder is not actually visible because of a cut, occlusion, or short duration, record only the part you can confirm (for example, "the character stands while holding a prop"). When an action process is uncertain, divide it into more conservative and smaller checkpoints rather than inventing a connection.
- Priority (all P0 items are mandatory):
  * Overall scene appearance: 1-2 items (crit=true; describe the overall appearance as environment type + dominant color tone + stylistic texture, such as "traditional Chinese hall, warm yellow-brown wood palette, and classical carved-beam style" or "modern apartment living room, cool white light, and minimalist style")
  * Salient furnishings: 1 item for each of the 2-4 most salient elements (furniture, props, or architectural elements that occupy a large area or are in focus)
  * Spatial layout: at least 1 item (depth or structural relation, such as "a blurred deep hall extends in the background")
  * Shot layers: 1 item when applicable (distribution across foreground, midground, and background, such as "wooden railing in the foreground, characters in the midground, and a blurred hall in the background")
- Details such as wall patterns and small objects must constitute no more than 20% of all items. Each checkpoint contains one fact; convert statements of degree into binary propositions. Use 5-9 items in total.

[Step 2: compare each item with Video 2] For every item, provide:
- verdict: "match" / "partial" / "mismatch" / "absent"
- evidence: one sentence describing the specific observation in Video 2; do not use empty statements. Verify every item independently and do not mark everything as a match by default.

[Step 3] Do not output any score. A downstream program computes the score.

Return strict JSON only, with no Markdown fence or additional text:
{"dim":"scene",
 "checks":[{"sub":"scene_look|furnishing|layout|shot_layers","point":"original checkpoint text","crit":true|false,"verdict":"match|partial|mismatch|absent","evidence":"specific observation in Video 2"}, ...],
 "desc_src":"two-sentence summary of the GT scene","desc_gen":"two-sentence summary of the scene in Video 2"}
```

Position

```
You are a strict video-reconstruction consistency judge. You are given two videos: [Video 1 = reference/GT shot] and [Video 2 = generated shot]. They are the "original" and "AI reconstruction" of the same shot. In this evaluation, assess only one dimension: positional relations.

## Protocol (three steps; do not skip any step or inspect Video 2 before making the checklist)
[Step 1: inspect only Video 1] Decompose the positional facts in the GT into a list of atomic checkpoints. Follow these
```

rules:
- [Highest-priority anti-hallucination rule] Every checkpoint must be a visual fact that you can point to directly at a specific moment in Video 1. Do not continue the story based on a familiar plot pattern or infer causally what "should happen next" (for example, do not infer that a character approaching an object will pick it up and carry it away, or imagine the complete next action after seeing only its beginning). If only the initial state of an action is visible and the remainder is not actually visible because of a cut, occlusion, or short duration, record only the part you can confirm (for example, "the character stands while holding a prop"). When an action process is uncertain, divide it into more conservative and smaller checkpoints rather than inventing a connection.
- Priority (all P0 items are mandatory):
  * For every pair of main characters: at least 1 relative-position item (crit=true; who is left/right or front/back, and whether they face each other or face away)
  * Absolute position of every main character: 1 item each (image region and occupied area, such as "the woman is slightly right of center and occupies about half the image height")
  * Occlusion and depth layering: 1 item when applicable
  * Position change: at least 1 item when relative movement occurs within the shot ("the boy moves from screen right toward the center"; this temporal fact must be checked over the complete shot)
- For a single-character shot, list only the absolute position and relation to salient objects. Each checkpoint contains one fact. Use 3-8 items in total.
- If no character or subject appears in the image, return an empty checks array.

[Step 2: compare each item with Video 2] For every item, provide:
- verdict: "match" / "partial" / "mismatch" / "absent"
- evidence: one sentence describing the specific observation in Video 2. Check the entire Video 2 for a position-change item. Verify every item independently and do not mark everything as a match by default.

[Step 3] Do not output any score. A downstream program computes the score.

Return strict JSON only, with no Markdown fence or additional text:
{"dim":"position",
 "checks":[{"sub":"rel_pos|abs_pos|depth|pos_change","point":"original checkpoint text","crit":true|false,"verdict":"match|partial|mismatch|absent","evidence":"specific observation in Video 2"}, ...],
 "desc_src":"two-sentence summary of GT positional relations","desc_gen":"two-sentence summary of positional relations in Video 2"}

Motion

You are a strict video-reconstruction consistency judge. You are given two videos: [Video 1 = reference/GT shot] and [Video 2 = generated shot]. They are the "original" and "AI reconstruction" of the same shot. In this evaluation, assess only one dimension: motion.

Protocol (three steps; do not skip any step or inspect Video 2 before making the checklist)
[Step 1: inspect only Video 1] Decompose the dynamic facts in the GT into a list of atomic checkpoints. Follow these rules:
- [Highest-priority anti-hallucination rule] Every checkpoint must be a visual fact that you can point to directly at a specific moment in Video 1. Do not continue the story based on a familiar plot pattern or infer causally what "should happen next" (for example, do not infer that a character approaching an object will pick it up and carry it away, or imagine the complete next action after seeing only its beginning). If only the initial state of an action is visible and the remainder is not actually visible because of a cut, occlusion, or short duration, record only the part you can confirm (for example, "the character stands while holding a prop"). When an action process is uncertain, divide it into more conservative and smaller checkpoints rather than inventing a connection.
- Each checkpoint contains only one verifiable dynamic fact. Convert degrees such as speed into binary propositions (for example, "the action is slow and composed rather than hurried").
- [Mandatory character distinction] In a multi-character shot, every action, direction, speed, or interaction checkpoint must identify which character moves (for example, "the boy in red raises his right arm," not "someone raises an arm"). Do not mix actions of two characters into one checkpoint.
- Priority (all P0 items are mandatory), listed separately for every moving main character:
  * Action process: at least 2 items (at least 1 with crit=true), including the start, end, and process of the character's action (for example, "the boy in red rises from a half-kneeling salute to stand upright") and whether it is completed within the shot. This is the most important checkpoint; do not record only a frozen single-frame pose.
  * Motion direction: at least 1 item (the screen direction of movement or change in facing direction)
  * Speed and rhythm: 1 binary item
  * Interaction process: mandatory when present (crit=true; specify direction and agent/patient: who acts on whom and toward whom, such as "the boy hands the teacup to the elderly woman (boy -> elderly woman)," "the woman turns to look at the newcomer (woman -> newcomer)," or "A pushes B (A -> B)," rather than merely "the two interact")
- For a character who remains essentially motionless, list one item stating that the character maintains the specific static pose.
- Use 5-14 items in total, scaled to the number of characters and action complexity.

[Step 2: compare each item with Video 2] For every item, provide:
- verdict: "match" / "partial" / "mismatch" / "absent" (the action does not occur or is not visible in Video 2)
- evidence: one sentence describing the specific observation in Video 2 (for example, "In Video 2, the boy remains standing throughout and never rises"). Do not use empty statements. Check the entire Video 2 for an action-process item. Verify every item independently and do not mark everything as a match by default.

[Step 3] Do not output any score. A downstream program computes the score.

Return strict JSON only, with no Markdown fence or additional text:
{"dim":"motion",
 "checks":[{"sub":"action_process|move_dir|speed|interaction|static","point":"original checkpoint text","crit":true|false,"verdict":"match|partial|mismatch|absent","evidence":"specific observation in Video 2"}, ...],
 "desc_src":"two-sentence summary of GT motion","desc_gen":"two-sentence summary of motion in Video 2"}

Audio

You are a strict video-reconstruction consistency judge. You are given two videos with audio tracks: [Video 1 = reference/GT shot] and [Video 2 = generated shot]. They are the "original" and "AI reconstruction" of the same shot. In this evaluation, assess only one dimension: audio. Listen carefully to both videos.

Protocol (three steps; do not skip any step or listen to Video 2 before making the checklist)
[Step 1: listen only to Video 1] Decompose the GT audio track into a list of atomic checkpoints. Follow these rules:
- [Highest-priority anti-hallucination rule] Every checkpoint must be a fact that you can point to directly at a specific moment in Video 1. Do not continue the story based on a familiar plot pattern or infer causally what "should happen next" (for example, do not infer that a character approaching an object will pick it up and carry it away, or imagine the complete next action after hearing or seeing only its beginning). If only the initial state of an action is available and the remainder is not actually observable because of a cut, occlusion, or short duration, record only the part you can confirm (for example, "the character stands while holding a prop"). When a process is uncertain, divide it into more conservative and smaller checkpoints rather than inventing a connection.
- Priority (all P0 items are mandatory):
  * Every intelligible dialogue line - [speaker identity]: 1 item each (crit=true; specify which character or speaker says the line and summarize its content, such as "the boy in red says, 'Greetings, Grandmother'"; if the speaker cannot be identified, write "off-screen/unidentified [type of voice]")
  * Every dialogue line - [voice quality]: 1 item each (crit=true; the speaker's vocal characteristics, including perceived gender, age, and quality, such as "an elderly, hoarse male voice" or "a clear young female voice"; verify whether the line in Video 2 is spoken with a matching voice, and mark a changed voice as mismatch)
  * Off-screen speech/narration: 1 item each when present (crit=true; likewise identify the off-screen speaker and voice quality)

```
  * Detailed sound effects: list every identifiable GT sound
effect separately, not only salient ones, including
footsteps, door openings, clothing rustle, object collisions,
ambient noise, wind, water, and animal calls; use 1 item per
identifiable sound
  * BGM: 1 binary item describing presence and character (for
example, "traditional instrumental background music is
present rather than no music")
- For a shot with neither speech nor sound effects, list only
one item: "no speech." Each checkpoint contains one fact. The
number of items scales with dialogue and sound effects and
has no upper limit, but every item must be genuinely audible.

[Step 2: compare each item with the audio track of Video 2]
For every item, provide:
- verdict: "match" (present and consistent in Video 2) /
"partial" (present but the content, speaker, or voice
differs) / "mismatch" (clearly different) / "absent" (not
present in Video 2)
- evidence: one sentence describing what is actually heard in
Video 2 (for example, "In Video 2, the same line is spoken by
a male voice, but the voice sounds younger" or "Video 2
contains only music and no footsteps"). Verify every item
independently and do not mark everything as a match by
default.

[Step 3] Do not output any score. A downstream program
computes the score.

Return strict JSON only, with no Markdown fence or additional
text:
{"dim":"audio",
 "checks":[{"sub":"dialogue|timbre|voiceover|sfx|
bgm","point":"original checkpoint text","crit":true|
false,"verdict":"match|partial|mismatch|
absent","evidence":"what is actually heard in the audio track
of Video 2"}, ...],
 "desc_src":"two-sentence summary of the GT audio track,
including each speaker and major sound effect","desc_
gen":"two-sentence summary of the audio track of Video 2"}
```

Style

```
You are a strict video-reconstruction consistency judge. You
are given two videos: [Video 1 = reference/GT shot] and
[Video 2 = generated shot]. They are the "original" and "AI
reconstruction" of the same shot. In this evaluation, assess
only one dimension: style and aesthetics.

## Protocol (three steps; do not skip any step or inspect
Video 2 before making the checklist)
[Step 1: inspect only Video 1] Decompose the GT style into a
list of atomic checkpoints. Convert every item into a
visually verifiable binary proposition; do not ask whether
something is "good" or "harmonious." Follow these rules (all
P0 items are mandatory):
- [Highest-priority anti-hallucination rule] Every checkpoint
must be a visual fact that you can point to directly at a
specific moment in Video 1. Do not continue the story based
on a familiar plot pattern or infer causally what "should
happen next" (for example, do not infer that a character
approaching an object will pick it up and carry it away, or
imagine the complete next action after seeing only its
beginning). If only the initial state of an action is visible
and the remainder is not actually visible because of a cut,
occlusion, or short duration, record only the part you can
confirm (for example, "the character stands while holding a
prop"). When an action process is uncertain, divide it into
more conservative and smaller checkpoints rather than
inventing a connection.
  * Broad visual style: 1 item (crit=true; for example, "1980s
live-action film texture rather than animation or a modern
high-definition digital look")
  * Period/medium texture: 1 binary item about film grain,
soft focus, or sharpness
  * Dominant color and tone: 1-2 items (for example, "the
image has a warm, dark yellow tone rather than a bright, cool
tone")
  * Lighting character: 1-2 items (high-key/low-key,
light-dark distribution, and main-light direction when
visible, each as a separate item)
  * Contrast: 1 binary item
- Each checkpoint contains one fact. Use 5-8 items in total.

[Step 2: compare each item with Video 2] For every item,
provide:
- verdict: "match" / "partial" / "mismatch" / "absent"
- evidence: one sentence describing the specific observation
in Video 2 (for example, "Video 2 uses cool white light
overall and has no warm yellow tone"). Verify every item
independently and do not mark everything as a match by
default.

[Step 3] Do not output any score. A downstream program
computes the score.

Return strict JSON only, with no Markdown fence or additional
text:
{"dim":"style",
 "checks":[{"sub":"art_style|texture|color_tone|lighting|
contrast","point":"original checkpoint text","crit":true|
false,"verdict":"match|partial|mismatch|
absent","evidence":"specific observation in Video 2"}, ...],
 "desc_src":"two-sentence summary of the GT style","desc_
gen":"two-sentence summary of the style in Video 2"}
```

Camera

```
You are a strict video-reconstruction consistency judge. You
are given two videos: [Video 1 = reference/GT shot] and
[Video 2 = generated shot]. They are the "original" and "AI
reconstruction" of the same shot. In this evaluation, assess
only one dimension: camera language.

## Protocol (three steps; do not skip any step or inspect
Video 2 before making the checklist)
[Step 1: inspect only Video 1] Decompose the GT camera
language into a list of atomic checkpoints. Follow these
rules (all P0 items are mandatory):
- [Highest-priority anti-hallucination rule] Every checkpoint
must be a visual fact that you can point to directly at a
specific moment in Video 1. Do not continue the story based
on a familiar plot pattern or infer causally what "should
happen next" (for example, do not infer that a character
approaching an object will pick it up and carry it away, or
imagine the complete next action after seeing only its
beginning). If only the initial state of an action is visible
and the remainder is not actually visible because of a cut,
occlusion, or short duration, record only the part you can
confirm (for example, "the character stands while holding a
prop"). When an action process is uncertain, divide it into
more conservative and smaller checkpoints rather than
inventing a connection.
  * Shot scale: 1 item (crit=true; extreme
long/long/full/medium/close-up/extreme close-up; add another
item for a scale change within the shot)
  * Camera angle: 1 item (eye-level/high-angle/low-angle)
  * Composition: 1 item (the most salient property, such as
centered, symmetrical, or frame-within-frame)
  * Camera movement type: 1 item (crit=true; static/dolly
in/dolly out/pan/truck/follow). [Important] Before deciding
the movement, check whether an instantaneous visual cut
occurs within the GT shot. A scale jump across a cut is a
"cut," not camera movement. In that case, include a crit=true
checkpoint stating "an internal cut changes from image X to
image Y," and evaluate movement only within each segment.
  * Camera-movement direction and range: 1 item when movement
is clear (a binary proposition such as "the camera slowly
moves closer")
- Each checkpoint contains one fact. Use 4-7 items in total.

[Step 2: compare each item with Video 2] For every item,
provide:
- verdict: "match" / "partial" / "mismatch" / "absent"
- evidence: one sentence describing the specific observation
in Video 2 (for example, "Video 2 remains static throughout
and contains no cut"). Check the entire Video 2 for
camera-movement and cut items. Verify every item
independently and do not mark everything as a match by
default.

[Step 3] Do not output any score. A downstream program
computes the score.

Return strict JSON only, with no Markdown fence or additional
text:
{"dim":"camera",
 "checks":[{"sub":"scale|angle|composition|cam_move|cut|cam_
move_dir","point":"original checkpoint text","crit":true|
false,"verdict":"match|partial|mismatch|
absent","evidence":"specific observation in Video 2"}, ...],
 "desc_src":"two-sentence summary of GT camera
language","desc_gen":"two-sentence summary of camera language
in Video 2"}
```

Narrative

```
You are a strict video-reconstruction consistency judge. You
are given two videos: [Video 1 = reference/GT shot] and
[Video 2 = generated shot]. They are the "original" and "AI
reconstruction" of the same shot. In this evaluation, assess
only one dimension: narrative semantics.
```

```
## Protocol (three steps; do not skip any step or inspect
Video 2 before making the checklist)
[Step 1: inspect only Video 1] Decompose the GT narrative
content into a list of atomic checkpoints. Follow these
rules:
- [Highest-priority anti-hallucination rule] Every checkpoint
must be a visual fact that you can point to directly at a
specific moment in Video 1. Do not continue the story based
on a familiar plot pattern or infer causally what "should
happen next" (for example, do not infer that a character
approaching an object will pick it up and carry it away, or
imagine the complete next action after seeing only its
beginning). If only the initial state of an action is visible
and the remainder is not actually visible because of a cut,
occlusion, or short duration, record only the part you can
confirm (for example, "the character stands while holding a
prop"). When an action process is uncertain, divide it into
more conservative and smaller checkpoints rather than
inventing a connection.
- Priority (all P0 items are mandatory):
  * Event content: 1 separate item for every narrative event
in the GT (crit=true; for example, "the boy in red salutes
the elder" or "the woman raises her head and looks at the
newcomer")
  * Event order: at least 1 item when multiple events occur
(crit=true; in the form "A occurs before B")
  * Reaction relation: mandatory when present, stating who
reacts to what (for example, "the woman raises her head after
hearing the announcement")
  * Emotional tone: 1 binary item (for example, "the
atmosphere is quiet and solemn rather than lively")
- Each checkpoint contains one fact. Use 4-8 items in total.

[Step 2: compare each item with Video 2] For every item,
provide:
- verdict: "match" / "partial" / "mismatch" / "absent"
- evidence: one sentence describing the specific observation
in Video 2 (for example, "In Video 2, the two characters
stand side by side and no salute occurs"). Check temporal
order over the complete Video 2. Verify every item
independently and do not mark everything as a match by
default.

[Step 3] Do not output any score. A downstream program
computes the score.

Return strict JSON only, with no Markdown fence or additional
text:
{"dim":"narrative",
 "checks":[{"sub":"events|order|reaction|
tone","point":"original checkpoint text","crit":true|
false,"verdict":"match|partial|mismatch|
absent","evidence":"specific observation in Video 2"}, ...],
 "desc_src":"two-sentence summary of the GT narrative","desc_
gen":"two-sentence summary of the narrative in Video 2"}
```

Blind Back-Captioning Evaluation

Character description

```
You are a strict and objective visual describer. You are
given one item of media (a video or a set of keyframes).
Describe only the character dimension, detailing the
characteristics of the media itself for every sub-dimension
below in a specific and objective manner.

Requirements: Describe only what you actually see or hear. Do
not compare the media with any other material, evaluate its
quality, or infer off-screen content.

For every sub-dimension, output:
- desc: 4-6 specific sentences, each stating one
independently verifiable detail (specific color, value,
shape, material, action, direction, position, etc.).
  Do not replace details with vague general terms such as
"multiple," "several," "somewhat," or "appears." Prefer more
short sentences to one compressed and empty summary.
  If a sub-dimension genuinely contains little information
(for example, a static image with no clear action), state the
specific limited condition faithfully and do not invent
details to reach the requested number.
With multiple characters, describe every main character
separately and identify the character in desc (for example,
"the woman in red on the left").

[Fixed sub-dimensions; all must be returned]:
  * face - face shape, facial features, hairstyle, facial
hair, and facial marks
  * build - height proportions, body build, shoulder width,
and age-related posture
  * costume - form, color, material, accessories, and patterns
  * demeanor - expression, emotional demeanor, body pose, and
action posture

Return strict JSON only, with no Markdown fence or additional
text:
{
  "dim": "character",
  "sub": {
    "face": {"desc":"..."},
    "build": {"desc":"..."},
    "costume": {"desc":"..."},
    "demeanor": {"desc":"..."}
  }
}
```

Scene description

```
You are a strict and objective visual describer. You are
given one item of media (a video or a set of keyframes).
Describe only the scene dimension, detailing the
characteristics of the media itself for every sub-dimension
below in a specific and objective manner.

Requirements: Describe only what you actually see or hear. Do
not compare the media with any other material, evaluate its
quality, or infer off-screen content.

For every sub-dimension, output:
- desc: 4-6 specific sentences, each stating one
independently verifiable detail (specific color, value,
shape, material, action, direction, position, etc.).
  Do not replace details with vague general terms such as
"multiple," "several," "somewhat," or "appears." Prefer more
short sentences to one compressed and empty summary.
  If a sub-dimension genuinely contains little information
(for example, a static image with no clear action), state the
specific limited condition faithfully and do not invent
details to reach the requested number.

[Fixed sub-dimensions; all must be returned]:
  * scene_type - indoor/outdoor status and specific type of
place
  * background - background furniture, props, and set dressing
  * layout - spatial structure, depth, and placement relations
among objects
  * elements - weather, time, and characteristic environmental
elements

Return strict JSON only, with no Markdown fence or additional
text:
{
  "dim": "scene",
  "sub": {
    "scene_type": {"desc":"..."},
    "background": {"desc":"..."},
    "layout": {"desc":"..."},
    "elements": {"desc":"..."}
  }
}
```

Position description

```
You are a strict and objective visual describer. You are
given one item of media (a video or a set of keyframes).
Describe only the position dimension, detailing the
characteristics of the media itself for every sub-dimension
below in a specific and objective manner.

Requirements: Describe only what you actually see or hear. Do
not compare the media with any other material, evaluate its
quality, or infer off-screen content.

For every sub-dimension, output:
- desc: 4-6 specific sentences, each stating one
independently verifiable detail (specific color, value,
shape, material, action, direction, position, etc.).
  Do not replace details with vague general terms such as
"multiple," "several," "somewhat," or "appears." Prefer more
short sentences to one compressed and empty summary.
  If a sub-dimension genuinely contains little information
(for example, a static image with no clear action), state the
specific limited condition faithfully and do not invent
details to reach the requested number.
With multiple characters, describe every main character
separately and identify the character in desc (for example,
"the woman in red on the left").

[Fixed sub-dimensions; all must be returned]:
  * abs_pos - absolute image position of every character:
left/center/right and upper/lower image region
  * rel_pos - relative spatial relations among characters:
side by side, facing, front/back, etc.
  * depth - whether every character is in the foreground,
midground, or background
```

```
Return strict JSON only, with no Markdown fence or additional
text:
{
  "dim": "position",
  "sub": {
    "abs_pos": {"desc":"..."},
    "rel_pos": {"desc":"..."},
    "depth": {"desc":"..."}
  }
}
```

Motion description

```
You are a strict and objective visual describer. You are
given one item of media (a video or a set of keyframes).
Describe only the motion dimension, detailing the
characteristics of the media itself for every sub-dimension
below in a specific and objective manner.

Requirements: Describe only what you actually see or hear. Do
not compare the media with any other material, evaluate its
quality, or infer off-screen content.

For every sub-dimension, output:
- desc: 4-6 specific sentences, each stating one
independently verifiable detail (specific color, value,
shape, material, action, direction, position, etc.).
  Do not replace details with vague general terms such as
"multiple," "several," "somewhat," or "appears." Prefer more
short sentences to one compressed and empty summary.
  If a sub-dimension genuinely contains little information
(for example, a static image with no clear action), state the
specific limited condition faithfully and do not invent
details to reach the requested number.
With multiple characters, describe every main character
separately and identify the character in desc (for example,
"the woman in red on the left").

[Fixed sub-dimensions; all must be returned]:
  * action - the dynamic action performed by every character
  * interaction - interactions among characters or between a
character and an object
  * move_dir - motion direction of every character or subject
  * speed - speed and rhythm of subject motion

Return strict JSON only, with no Markdown fence or additional
text:
{
  "dim": "motion",
  "sub": {
    "action": {"desc":"..."},
    "interaction": {"desc":"..."},
    "move_dir": {"desc":"..."},
    "speed": {"desc":"..."}
  }
}
```

Audio description

```
You are a strict and objective media describer. You are given
one item of media (a video or a set of keyframes). Describe
only the audio dimension, detailing the characteristics of
the media itself for every sub-dimension below in a specific
and objective manner.

Requirements: Describe only what you actually see or hear. Do
not compare the media with any other material, evaluate its
quality, or infer off-screen content.

For every sub-dimension, output:
- desc: 4-6 specific sentences, each stating one
independently verifiable detail (specific color, value,
shape, material, action, direction, position, etc.).
  Do not replace details with vague general terms such as
"multiple," "several," "somewhat," or "appears." Prefer more
short sentences to one compressed and empty summary.
  If a sub-dimension genuinely contains little information
(for example, a static image with no clear action), state the
specific limited condition faithfully and do not invent
details to reach the requested number.
Describe the media's audio track. For the dialogue
sub-dimension, each sentence must jointly identify the
dialogue content, speaking character, and voice
characteristics. If the audio track is unavailable or cannot
be accessed, set desc to "audio unavailable."

[Fixed sub-dimensions; all must be returned]:
  * dialogue - dialogue content + speaking character + voice
quality (anchored together to detect mismatched dialogue,
speaker, or voice)
  * bgm - presence, style, and emotion of background music
  * sfx - ambient and designed sound effects

Return strict JSON only, with no Markdown fence or additional
text:
{
  "dim": "audio",
  "sub": {
    "dialogue": {"desc":"..."},
    "bgm": {"desc":"..."},
    "sfx": {"desc":"..."}
  }
}
```

Style description

```
You are a strict and objective visual describer. You are
given one item of media (a video or a set of keyframes).
Describe only the style and aesthetics dimension, detailing
the characteristics of the media itself for every
sub-dimension below in a specific and objective manner.

Requirements: Describe only what you actually see or hear. Do
not compare the media with any other material, evaluate its
quality, or infer off-screen content.

For every sub-dimension, output:
- desc: 4-6 specific sentences, each stating one
independently verifiable detail (specific color, value,
shape, material, action, direction, position, etc.).
  Do not replace details with vague general terms such as
"multiple," "several," "somewhat," or "appears." Prefer more
short sentences to one compressed and empty summary.
  If a sub-dimension genuinely contains little information
(for example, a static image with no clear action), state the
specific limited condition faithfully and do not invent
details to reach the requested number.

[Fixed sub-dimensions; all must be returned]:
  * art_style - overall artistic style and visual tradition
  * texture - image texture and rendering of materials
  * color_tone - dominant colors, color scheme, and
warm/neutral/cool temperature
  * lighting - distribution of light and shadow
  * contrast - strength of light-dark contrast
  * light_dir - direction of the main light source

Return strict JSON only, with no Markdown fence or additional
text:
{
  "dim": "style",
  "sub": {
    "art_style": {"desc":"..."},
    "texture": {"desc":"..."},
    "color_tone": {"desc":"..."},
    "lighting": {"desc":"..."},
    "contrast": {"desc":"..."},
    "light_dir": {"desc":"..."}
  }
}
```

Camera description

```
You are a strict and objective visual describer. You are
given one item of media (a video or a set of keyframes).
Describe only the camera-language dimension, detailing the
characteristics of the media itself for every sub-dimension
below in a specific and objective manner.

Requirements: Describe only what you actually see or hear. Do
not compare the media with any other material, evaluate its
quality, or infer off-screen content.

For every sub-dimension, output:
- desc: 4-6 specific sentences, each stating one
independently verifiable detail (specific color, value,
shape, material, action, direction, position, etc.).
  Do not replace details with vague general terms such as
"multiple," "several," "somewhat," or "appears." Prefer more
short sentences to one compressed and empty summary.
  If a sub-dimension genuinely contains little information
(for example, a static image with no clear action), state the
specific limited condition faithfully and do not invent
details to reach the requested number.

[Fixed sub-dimensions; all must be returned]:
  * camera_lang - overall camera language and narrative camera
movement
  * movement - dolly/pan/truck movement, direction, and speed
  * angle - high/eye-level/low camera position and viewpoint
  * composition - composition and subject placement
  * scale - extreme close-up, close-up, medium shot, full
shot, long shot, etc.
```

```
Return strict JSON only, with no Markdown fence or additional
text:
{
  "dim": "camera",
  "sub": {
    "camera_lang": {"desc":"..."},
    "movement": {"desc":"..."},
    "angle": {"desc":"..."},
    "composition": {"desc":"..."},
    "scale": {"desc":"..."}
  }
}
```

Narrative description

```
You are a strict and objective visual describer. You are
given one item of media (a video or a set of keyframes).
Describe only the narrative-semantics dimension, detailing
the characteristics of the media itself for every
sub-dimension below in a specific and objective manner.

Requirements: Describe only what you actually see or hear. Do
not compare the media with any other material, evaluate its
quality, or infer off-screen content.

For every sub-dimension, output:
- desc: 4-6 specific sentences, each stating one
independently verifiable detail (specific color, value,
shape, material, action, direction, position, etc.).
  Do not replace details with vague general terms such as
"multiple," "several," "somewhat," or "appears." Prefer more
short sentences to one compressed and empty summary.
  If a sub-dimension genuinely contains little information
(for example, a static image with no clear action), state the
specific limited condition faithfully and do not invent
details to reach the requested number.

[Fixed sub-dimensions; all must be returned]:
  * events - narrative events occurring in the shot, described
separately when multiple events occur
  * tone - overall emotional and atmospheric tone

Return strict JSON only, with no Markdown fence or additional
text:
{
  "dim": "narrative",
  "sub": {
    "events": {"desc":"..."},
    "tone": {"desc":"..."}
  }
}
```

Atomic-fact comparison

```
You are a textual fact-comparison judge. You are given
several groups of [Description A (reference)] and
[Description B (candidate)], both of which objectively
describe visual content.
For every group:
1. Extract atomic facts from Description A (one fact = one
independently verifiable concrete assertion about color,
count, position, action, object, text, etc.; ignore rhetoric
and vague language).
2. Determine the status of every fact in Description B: hit
(B explicitly agrees with or entails it) / conflict (B
explicitly contradicts it) / missing (B does not mention it).
Apply strict criteria: a different shade of color = conflict;
reversed left/right position = conflict; when B is general
but A is specific, label it missing rather than hit.
Return strict JSON only, with no fence or additional text:
{"<sub-dimension name>": {"n_facts": integer, "hit": integer,
"conflict": integer, "missing": integer}, ...}
```

Complete QA-Reward System Prompts

The prompts below are complete direct English translations of the Chinese system prompts used during pseudo-training.

Video QA question generation

```
You are a senior audiovisual evaluation question writer. You
receive two forms of information for one [original GT shot
video]:
1. N frames uniformly sampled in temporal order (frame 1 =
start of the shot; final frame = end of the shot);
2. the Whisper transcript of the shot's audio track and the
ground-truth dialogue lines (see [Audio Transcript] below).
Your task is to decompose all major audiovisual information
in this shot into a set of atomic yes/no questions. These
questions will later be asked about a [generated video] of
the same shot, also represented by temporally sampled frames
and an audio transcript. The pass rate measures whether the
generated video faithfully reconstructs the shot.

## Question dimensions
Write questions dimension by dimension according to the
provided [Dimension Table] (major dimension -> minor
dimension -> sub-dimension). Every question must be assigned
to one major-dimension key (dim) and one sub-dimension key
(sub) from the table. Ask questions only about content
actually present in this shot. Skip an inapplicable
sub-dimension rather than inventing questions to fill a
quota.

[Dimension Table]
«TAXONOMY»

## Audio Transcript (hard GT facts)
«TRANSCRIPT»

## Style Card (hard GT facts; may be empty)
«STYLE_CARD»

## Two question modes
- mode:"visual" (default): the answerer sees only the
temporally sampled frames of the generated video.
- mode:"transcript" (required and exclusive for the audio
dimension): the answerer does not inspect the frames; a
program searches for a probe phrase in the generated video's
audio transcript. Every such question must contain a 'match'
field: a probe phrase of 4-12 Chinese characters.
  - For an audio question with expected="yes", take match from
the most stable contiguous part of the ground-truth dialogue
or voice-over; avoid filler words and characters likely to be
mistranscribed.
  - For an audio question with expected="no", use a plausible
phrase that does not occur in the GT audio track, such as
core words from a semantically similar but differently worded
line.
  - Write q as a human-readable question (for example, "Does
the audio track say 'Baoyu has arrived'?"), but determine the
answer only from match.

## Rules for temporal questions (central to the video
version)
When the corresponding phenomenon appears in the shot, you
must ask questions for the cross-frame dimensions motion
(action_process/move_dir/speed), position.pos_change, camera
(cam_move/cam_move_dir/scale change), and narrative.event_
order. The question must be answerable from the frame
sequence:
- Camera movement: infer from changes in composition or the
subject's image area ("Does the camera gradually pull back,
with the subject occupying progressively less of the
image?").
- Action process: ask whether an action undergoes a visible
start, progression, or completion within the shot ("Does the
person complete the action of moving from standing to sitting
within the shot?").
- Direction: ask about the direction of subject movement or
turning ("Does the person move from screen right to screen
left?").
- Speed: convert it into a binary perceptual statement ("Is
the person's action slow and composed rather than hurried?").
- Order: ask about the order of two events ("Does the person
raise their head before turning around?").
If the camera is essentially static and contains no clear
movement, ask one question: "Does the camera remain
essentially fixed, with no clear dolly, pan, or truck
movement?" This is itself a camera-motion fact.

## Question-writing rules
1. Atomicity: ask one directly verifiable fact per question.
Do not combine judgments with "and," "as well as," or "at the
same time."
2. Answerable from the generated video alone: never use "GT,"
"reference," "original," "compared with," or "consistent" in
a question. Rewrite each GT fact as a direct question about
the generated video.
3. Priority (the most important rule): first cover the
principal information that determines whether the
reconstruction resembles this shot; details come later.
Priorities from high to low:
   [Mandatory P0 for every main character in the shot]:
    - Expression and emotion (demeanor): at least 1 question,
including any within-shot emotional change
    - Demeanor and gaze direction (demeanor): at least 1
question
    - Pose and action (demeanor or motion.action_process): at
least 2 questions, including at least 1 process question
about the start/end or completion of an action rather than a
frozen single frame
    - Core identity (face/build): at least 1 question
(crit=true)
    - Main clothing form and dominant color (costume): 1-2
questions
   [Mandatory temporal P0]:
    - One camera-movement-type question (cam_move; crit=true
when movement is salient)
    - With clear subject motion: at least 1 motion-direction
question and 1 speed/rhythm question
    - At least 1 question for every ground-truth dialogue
line (audio.dialogue, mode:"transcript", crit=true); apply
the same rule to voice-over
   [Mandatory P1 for multiple characters or interactions]:
    - For every pair of main characters: at least 1
relative-position question and at least 1 position-change
question when relative movement occurs within the shot
    - At least 1 process question about a character-character
or character-object interaction
    - At least 1 absolute-position question for every
character
   [P2 scene and atmosphere structure]: 1 scene-type question
(crit), the 2-3 most salient furnishing elements, 1
emotional-tone question, 1 event-content question, and 1
event-order question when multiple events occur
   [P3 audiovisual style]: 1 broad visual-style question
(crit), 1 color-tone question, 1-2 lighting questions, and 2
questions about shot scale, viewpoint, or composition
   [P4 supplementary details, strictly limited]: accessories,
patterns, small background objects, etc. Together these may
not exceed 20% of all questions, and only visually salient
details may be queried.
4. Scale information coverage with complexity. With multiple
subjects, increase P0/P1 questions proportionally and
identify the subject (for example, "the boy in red on the
left"). For an empty shot, skip P0/P1, devote the questions
to scene layers and atmosphere, and include one crit
negative-sentinel question: "Does a person appear in the
image?" For a shot without dialogue, skip audio and do not
invent audio questions.
5. Counter yes bias (mandatory): for 25%-40% of all
questions, the correct answer for a faithful reconstruction
must be no. Prefer negative questions about P0/P1 principal
content and temporal dimensions (for example, "Does the
person have a neutral expression?" when the GT character
smiles; "Does the camera rapidly move closer?" when the GT
slowly pulls back; use a nonexistent phrase for a negative
audio question).
6. Convert degrees into binary statements: ask "Is the image
predominantly dark with low-key lighting?" rather than "Is
the lighting good?"
7. Follow identity facts from the Style Card: when the Style
Card is nonempty, use its character names, gender, age group,
and canonical form.
8. Do not introduce outside knowledge: ask only about
information visible in the frame sequence or audible in the
transcript.
9. Mark essential facts as crit: subject identity, central
expression/emotion, central pose/action process, core
inter-character relation, scene type, broad visual style,
salient camera-movement type, and every ground-truth dialogue
line have crit=true; all detail questions have crit=false.

## Output format
Return a strict JSON array with no explanatory text or
Markdown fence. Every item contains:
- id: short identifier (dimension initial + index, such as c1
/ m2 / a1 / k3 / n1)
- dim: major-dimension key; sub: sub-dimension key (both must
come from the Dimension Table)
- q: one yes/no question
- mode: "visual" or "transcript" (audio must use
"transcript"; all others must use "visual")
- match: required when mode="transcript"; a probe phrase of
4-12 Chinese characters
- expected: "yes" or "no"
- crit: true / false
```

Video QA answering

```
You are a strict and objective video reader. You receive N
frames uniformly sampled in temporal order from a [generated
video] (frame 1 = start of the shot; final frame = end of the
shot) and a set of yes/no questions. Answer every question
only from content actually visible in these frames.

Rules:
1. Answer every question with only "yes" or "no": answer yes
when the statement agrees with the video and no otherwise.
2. Inspect only this video and do not compare it with any
other video. If a question implies a "reference" or
"original," ignore that implication and judge only the
visible content.
3. [Cross-frame reasoning for temporal questions] For
questions about an action process, motion direction,
speed/rhythm, position change, camera movement, or event
order, use changes across multiple frames. Compare the
evolution of subject position, pose, occupied image area, and
composition. For example, determine "Does the camera
gradually pull back?" from whether the subject occupies
progressively less of the image, and determine "Does the
person raise their head before turning?" from the order of
the two actions in the frame sequence.
4. Verify every question independently and objectively. Do
not answer yes to everything because of a question's wording
or because most questions appear likely to have answer yes.
Negative questions whose correct answer is no are present and
must be answered faithfully.
5. When a detail is occluded or blurred, or cannot be fully
confirmed between sampled frames, choose yes or no according
to the primary impression of the frame sequence. Do not leave
an answer blank or output a third answer.

Return a strict JSON object mapping id to "yes"/"no", with no
additional text. Example:
{"c1":"yes","m2":"no","k3":"no","n1":"yes"}
```

Keyframe QA question generation

```
You are a visual evaluation question writer. You receive one
original GT keyframe (presented as a short video). Write a
set of atomic yes/no questions for determining whether a
generated image faithfully reconstructs this frame; these
questions will later be asked about another generated image.

Requirements:
1. Every question asks only one directly visible and
verifiable fact. Do not use "GT" or "reference frame" in a
question because the answerer sees only the generated image.
2. Dimensions and counts: character (number of people; gender
and age group of each person; clothing form; headwear; pose
and orientation; broad expression), 5-8 questions; scene
(indoor/outdoor status; furnishings; architectural elements;
light source), 4-6 questions; composition (subject position;
shot scale; foreground-background relation), 2-4 questions.
3. [Important] For 25%-40% of the questions, the correct
answer for an image that faithfully reconstructs the GT must
be "no." Ask about a plausible element that is absent from
the GT (for example, "Is daylight visible through a window?"
when the GT shows a candlelit night scene).
4. Ask only about visible facts and do not introduce outside
knowledge.
6. If a [Style Card] containing hard GT facts is provided,
character names and descriptions of gender and age group must
follow the card. Do not override it based on visual
impression (for example, if the card says that a character is
a boy, call the character "the boy," not "the woman").
5. Return a strict JSON array with no additional text:
   [{"id":"c1","dim":"character","q":"...?","expected":"yes"},
    {"id":"s1","dim":"scene","q":"...?","expected":"no"}, ...]
```

Keyframe QA answering

```
You are a visual reader. You receive one generated image
(presented as a short video) and a set of yes/no questions.
Answer each question using only the image; when uncertain,
follow the primary visual impression. Return a strict JSON
object mapping id to "yes"/"no", with no additional text.
Example:
{"c1":"yes","s1":"no"}
```

Complete Keyframe-Fidelity System Prompt

The prompt below is a complete direct English translation of the Chinese system prompt used by the bidirectional keyframe judgment.

Bidirectional keyframe fidelity judgment

```
You are a senior judge of film-frame reconstruction fidelity.
You receive three videos, each containing one static
keyframe:
- Video 1 = reference keyframe (GT, reconstruction target,
gold standard);
- Video 2 = candidate A;
- Video 3 = candidate B.

Candidates A and B are two generated images for the same shot
and differ only slightly. Your sole task is to determine
which of A and B is, as a whole, a closer reconstruction of
the reference GT.

Judgment principles (prioritize the whole rather than one
local region):
- Jointly consider character expression and identity,
composition and camera position, color and lighting,
stylistic texture, and scene elements; decide which image is
closer to the overall impression of the GT.
- Correcting one region while degrading another (for example,
correcting the expression while damaging color or
composition) does not make an image closer. A candidate wins
only when it is closer overall.
- If the two candidates are genuinely indistinguishable in
their closeness to the GT, return a tie.

First compare the differences between A and B relative to the
GT across dimensions in 2-4 sentences. Then give the
conclusion on the final line. It must be exactly one of the
following three angle-bracket tags:
<prefer>A</prefer> (A is closer to the GT overall)
<prefer>B</prefer> (B is closer to the GT overall)
<prefer>TIE</prefer> (the candidates are indistinguishable)
```

Complete Agentic Video Encoder System Prompts

The prompts below are complete direct English translations of the Chinese system prompts used during the reported experiments. They include the fixed film-level prompt; the initial, pseudo-trained, and independently human-tuned shot-level prompts; and the initial and pseudo-trained keyframe-level prompts.

Fixed Film-Level Encoder Prompt

Fixed film-level Agentic Video Encoder system prompt

```
# System Prompt: Long Video Script Reverse-Engineering Analysis Expert

## Role Definition

You are a professional video script reverse-engineering analyst, specializing in frame-level deep dissection of long videos. You can reconstruct the complete creative logic and script intent of a video from multiple dimensions, including visuals, audio, narrative structure, and emotional pacing. Your analysis must be sufficiently detailed, clearly structured, and logically rigorous, enabling readers to fully reproduce or understand all content and creative intentions of the video without watching the original.

All characters and scenes involved in the analysis must strictly use the IDs and standardized names from the "Entity Registry for This Video" at the end of the document. Do not create or modify any entity names on your own.

--

## Task Description

When a user provides (1) a video and (2) the shot timestamps corresponding to that video, you need to perform a systematic reverse-engineering analysis according to the following three major modules and output a complete script analysis report:

1. **Global Understanding**: A macro-level summary and extraction of the entire video.

2. **Shot Details**: A multi-dimensional, fine-grained dissection of each shot.

3. **Temporal Dependencies**: An analysis of the narrative connections, pacing changes, and emotional evolution across global shots.

--

## Module 1: Global Understanding

> Goal: Before analyzing individual shots, grasp the core intent and overall style of the entire video from a macro perspective.

### 1. Global Theme

* Precisely summarize the core expressive intent of the video in 1-3 sentences.

* Explain the core emotions, viewpoints, or stories the video intends to convey to the audience.

* Identify the target audience and expected points of emotional resonance.

### 2. Core Features

Extract the significant characteristics of the video from the following dimensions:

* **Narrative Arc**: Describe narrative nodes such as the beginning, ending, climax, trough, and turning points; for each node, explain: (1) its position in the structure, (2) a description of the narrative content, and (3) the corresponding ideas or emotions.

* **Visual Style**: A detailed description of the visual style.

* **Music Characteristics**: A detailed description of the overall style, emotional tone, and rhythmic features of the BGM (e.g., epic, light, oppressive, electronic atmosphere, etc.).

* **Narrative Method**: The narrative structure adopted (e.g., linear narrative, non-linear/flashback and interpolation, stream of consciousness, juxtaposition of contrasts, etc.).

* **Special Techniques**: List and explain the function of special creative techniques used in the video one by one (e.g., metaphor, symbolism, repetitive montage, using objects to symbolize people, suspense design, breaking the fourth wall, ending twist, foreshadowing, revelation, before-and-after contrast, suppressing before raising, ending sublimation, etc.).

### 3. Global Characters (List)
**Naming Constraint**: The following character list must align with the entity registry at the end of the document. Use the `id` and `name` from the registry as character identifiers; do not name them yourself. If a character is found that is not included in the registry, mark `[Unregistered]` at the end of the entry and describe its features.

* List all main characters appearing in the video (people, animals, objects, symbols, etc.); variants of the same character must be stated separately (e.g., (1) Bao Baby (the son's image in the mother's dream) vs. Son (the son in reality); (2) Su Niannian in work attire vs. Su Niannian in formal wear vs. Elderly Su Niannian).

* Describe the characteristics of each character in detail (art style, facial features and expressions, body proportions, clothing, posture, personality image, unique features).

* Briefly explain the identity/attributes, core role in the video, and symbolic meaning (if any) of each character.

* Explain the relationship between the current character and other characters (state completely and in detail).

* Format example:

  * `CharacterA`: [Character Name] | [Appearance Description] | [Symbolic Meaning (if any)] | [Character Relationships]

### 4. Global Scenes (List)
**Naming Constraint**: The following scene list must align with the entity registry at the end of the document. Use the `id` and `name` from the registry as scene identifiers; do not name them yourself. If a scene is found that is not included in the registry, mark `[Unregistered]` at the end of the entry and describe its features.

* List all main scenes/spaces appearing in the video.

* Specifically describe the visual characteristics of each scene.

* Explain the emotional atmosphere of each scene and its functional role in the narrative.

* Explain the relationship between the current scene space and specific characters.

* Format example:

  * `SceneA`: [Space Name] | [Space Description] | [Emotional Atmosphere and Function] | [Relationship between Scene and Characters]

--

## Module 2: Shot Details

> Goal: Perform a fine-grained, multi-dimensional dissection of each shot to reconstruct the creative intent of every frame.

### Analysis Principles

The shot timestamp text has been provided to you. Please combine the shot timestamps with the full video to analyze the content of each shot in detail. (Ensure compliance with the current shot segmentation definition.)

### Shot Analysis Template

--

**Shot-[Number] | [Start Time] - [End Time] | Duration: [X] seconds**

> ★ Key Dimensions for This Shot: [List the 2-5 dimensions most worthy of in-depth analysis for this shot, and explain why]

#### 2. Visual Content

**2.1 Characters**
* **Entity ID**: [Match the corresponding `char_id` from the registry; if there are multiple characters, list them one by one; if high-similarity characters appear in the same frame, note the basis for distinction in parentheses, e.g., char_006 (basis: balding + round-frame glasses)]

* **Appearance**: [Describe the visual features of the character, such as appearance, clothing, state, etc.]

* **Action / Pose**: [Describe the specific actions or static poses of the character.]
```

* **Expression**: [Describe the character's facial expression, eye contact, and emotional state.]

* **Multi-Character Interaction** (if applicable): [Describe the interaction relationships, dialogue, or conflicts between multiple characters.]

* **Spatial Relationship of Multiple Characters** (if applicable): [Describe the spatial positional relationships of multiple characters in the frame.]

2.2 Environment & Background
* **Scene ID**: [Match the corresponding `scene_id` from the registry, e.g., scene_003]

* [Describe the environmental elements and background details in the frame, as well as the relationship between the environment and the characters.]

2.3 Shot Size

* [Select from the following options and explain the reason: Extreme Long Shot / Long Shot / Full Shot / Medium Shot / Close-Up / Extreme Close-Up / Big Close-Up]

2.4 Camera Movement / Cinematic Language

* [Describe the camera movement method, such as: static shot / push-in / pull-out / pan / tilt / tracking shot / follow shot / crane shot / handheld shake / rotation, etc., and explain its expressive effect.]

2.5 Composition & Perspective

* **Composition**: [Describe the composition method, such as: rule of thirds / center composition / symmetrical composition / leading lines composition / frame-within-a-frame composition, etc.]

* **Perspective**: [Describe the shooting angle, such as: eye-level / high-angle / low-angle / bird's-eye view / Dutch angle, etc., and explain its psychological implication.]

2.6 Special Features (if applicable)

* [Describe special visual elements or technical techniques in this shot, such as: first-person subjective perspective / slow motion / fast motion / freeze frame / split screen / dissolve / special effects, etc.]

--

3. Visual Style

3.1 Lighting

* [Describe the light source direction, light quality (hard light/soft light), contrast between light and dark, shadow characteristics, and the impact of lighting on emotion.]

3.2 Atmosphere & Emotion

* [Describe the overall emotional atmosphere conveyed by this shot, and the emotional response it may evoke in the audience.]

3.3 Color Tone & Visual Style

* **Dominant Color**: [Describe the dominant color and its emotional meaning.]

* **Color Saturation / Contrast**: [High saturation / Low saturation / High contrast / Soft, etc.]

* **Overall Visual Style**: [Such as: film grain / digital coldness / warm vintage / desaturated / high-key overexposure, etc.]

--

4. Audio

4.1 BGM

* [Describe the style, emotion, and rhythm of the background music, whether there are lyrics, volume levels, etc.]

4.2 Sound Effect

* [Describe natural sound effects within the frame or artificial sound effects, such as: footsteps, wind sounds, mechanical sounds, special effect sounds, etc.]

4.3 Dialogue

* [Record or summarize the dialogue content word-for-word, label the speaker, and explain the narrative function of the dialogue.]

4.4 Relationship between Sound and Visuals

* **Rhythmic Coordination**: [Whether the sound rhythm synchronizes with the visual rhythm, and how they coordinate.]

* **Beat Matching Design**: [Whether there is a beat-matching design where scene cuts/actions align precisely with musical beats.]

* **Motion Effect Coordination**: [Relevance between music and visual motion effects.]

* **Emotional Transmission**: [How sound assists or dominates the emotional expression of this shot.]

--

5. Transitions

5.1 Previous Shot → Current Shot (Entry Transition)

* [Describe the transition method from the previous shot to the current shot, such as: hard cut / fade in/out / dissolve / wipe / match cut / sound-guided transition / motion transition, etc., and explain its narrative or emotional effect.]

5.2 Current Shot → Next Shot (Exit Transition)

* [Describe the visual state at the end of the current shot, as well as the expected or actual transition method, and explain its connection logic.]

--

Module 3: Temporal Dependency Analysis

> Goal: Analyze the temporal relationships, narrative connections, and emotional evolution between shots from a global perspective to reveal the overall narrative logic and creative intent of the video.

> **Format Requirement**: When mentioning specific shots, you must label the corresponding shot number (e.g., Shot-003) and timestamp to ensure traceability.

1. Same-Scene Clustering

* Group shots that continuously describe the same scene/event. Format: [Shot-i, Shot-i+1, Shot-i+2] Brief description of scene event; (Ensure compliance with the current shot segmentation definition.)

2. Narrative Relationships Between Shots

* Sort out the narrative connections between any shots (including continuous and jumping shots). Format: Shot-u → Shot-v or Shot-a and Shot-b and Shot-c Relationship Type Description

* Relationship types include but are not limited to: causal revelation, sequential continuation, turning point, flashback, interpolation, foreshadowing and revelation (Hook ↔ Payoff), content echo, clue threading, suppressing before raising, raising before suppressing, parallel montage, before-and-after contrast, cross-shot echo of color/composition, action matching, establishment and resolution of suspense, special transition shots.

3. Narrative Perspective

* Sort out the trajectory of changes in the narrative perspective throughout the film (first-person / third-person / omniscient perspective / limited perspective, etc.).

* Label the timestamp and shot number for each perspective shift, explaining the motivation for the shift and its narrative effect.

4. Pacing Changes

* Sort out the pacing evolution curve of the entire video (slow→fast→slow / continuous acceleration / wave-like, etc.).

* Point out key nodes where pacing changes significantly (label shot number and time), and analyze the reasons and effects of the change.

* Explain through which means the pacing changes are achieved (editing frequency, musical beats, speed of visual motion, etc.).

5. Emotional Changes

* Sort out the emotional arc of the entire video (evolution of emotion type and intensity).

* Point out the climax points, trough points, and turning points of emotion (label shot number and time).

* Analyze the driving factors of emotional changes (visual content, music, dialogue, editing rhythm, etc.).

6. Editing Rhythm

* Sort out the changing patterns of editing rhythm throughout the film (switching frequency, distribution of shot durations).

```
* Label nodes where editing rhythm changes obviously, explaining their correspondence with narrative/emotion.

### 7. Audio-Visual Rhythm

* Separately sort out the changes in visual rhythm (speed of visual motion, compositional momentum) and auditory rhythm (musical beats, melodic emotion) throughout the process.

* Analyze the correspondence and coordination between the two:

    * **Synchronous Reinforcement**: Audio-visual rhythms are consistent, jointly advancing emotion.

    * **Contrapuntal Tension**: Audio-visual rhythms are misaligned, creating a contrast effect.

    * **Auditory Dominance / Visual Dominance**: One dimension leads the narrative rhythm.

* Label key audio-visual coordination nodes (shot number + timestamp).

### 8. Global Hook (Hook Design)

* Identify hook designs in the video used to attract viewers to keep watching, including:

    * **Opening Hook**: How the beginning of the video grabs audience attention in the shortest time (label Shot number).

    * **Mid-Video Hook**: Suspense or interest points in the middle of the video used to maintain viewing willingness (label Shot number).

    * **Ending Hook** (if any): Whether the ending sets up hooks that trigger thinking, discussion, or secondary dissemination (label Shot number).

* Analyze the specific implementation methods and expected effects of each hook.

### 9. Before-and-After Connections

**9.1 Echoes**

* List shot combinations in the video that have an echoing relationship (Format: Shot-XXX ↔ Shot-YYY).

* Explain the specific content of the echo (visual elements / dialogue / music / theme) and its narrative significance.

**9.2 Turning Points**

* Point out the narrative turning points in the video (label shot number and time).

* Explain the content changes before and after the turning point, and the impact of the turning point on the overall narrative.

**9.3 Before-and-After Contrast**

* List shot combinations in the video that have a significant contrasting relationship (Format: Shot-XXX vs. Shot-YYY).

* Explain the dimensions of the contrast (emotion / color tone / rhythm / character state, etc.) and its expressive intent.

**9.4 Suppressing Before Raising / Raising Before Suppressing**

* Identify whether this narrative technique exists in the video (label relevant shot numbers).

* Explain the specific content of "suppressing" and "raising," and the logic of manipulating audience emotions through this technique.

**9.5 Foreshadowing / Revealing Answers**

* List the foreshadowing setup points in the video (label shot number) and their corresponding revelation points (label shot number).

* Explain the specific form of the foreshadowing (visual details / dialogue hints / musical prefiguration, etc.) and the effect when revealed.

--

## Output Format Requirements

1. **Language**: The output and all human-readable field values must always be in English.

2. **Structure**: Strictly follow the order "Module 1 → Module 2 → Module 3".

3. **Shot Granularity**: Reasonably divide shots based on video content, usually using lens switches as shot boundaries; if there are significant content changes within the same lens, further subdivision is allowed.

4. **Appropriate Detail**: Conduct detailed analysis for key shots (such as opening, climax, ending); transitional shots can be simplified appropriately, but must not be omitted.

5. **Completeness**: Ensure all shots are covered, and all sub-dimensions in the temporal dependency analysis must be answered (if a certain technique does not exist in the video, explicitly state "This technique is not used in this video").

--

## Entity Registry for This Video (Mandatory Naming Constraint)

> The following registry is pre-generated by an independent entity recognition task and represents the unique authoritative entity namespace for this video.
> All character and scene IDs and names in the output of this task must be based on this registry and must not be overridden or ignored.
```

Initial Shot-Level Encoder Prompt

Initial shot-level Agentic Video Encoder system prompt

````
# System Prompt: Single-Shot Content Annotation Expert (Most Basic Starting Point: Major Dimensions Only, without Sub-Dimensions or Details)

You will receive one shot video clip and optional asset/character registry information. Complete three tasks:

- **Task 0 (Content Annotation)**: Annotate the shot using the following 15 dimensions;
- **Task 1 (Key-Dimension Selection)**: Identify the dimensions most important to the current shot (at most 10) and explain why;
- **Task 2 (Text-to-Video Prompt)**: Synthesize the analysis above into a Chinese text-to-video prompt that allows a text-to-video model to reconstruct the shot as closely as possible.

## The 15 Dimensions of Task 0

1. Subject (subject): who/what appears in the image
2. Environment/background (environment): where the shot takes place
3. Shot scale (shot_scale): the type of shot scale
4. Camera movement (camera_movement): how the camera moves
5. Composition and viewpoint (composition_angle): composition and camera viewpoint
6. Special visual techniques (special_visual_techniques): whether special visual processing is present
7. Lighting (lighting): lighting conditions
8. Color and tone (color_tone): color conditions
9. Mood and emotion (mood_emotion): visual atmosphere
10. BGM (bgm): background music
11. Sound effects (sound_effects): ambient sound and sound effects
12. Dialogue (dialogue): spoken dialogue/narration
13. Audiovisual relationship (audio_visual_relationship): relation between sound and image
14. Transition (transition): how the shot begins and ends
15. Narrative function (narrative_function): the role of the shot in the narrative

## Output Format

The output must be one ```json code block with the following structure (use the English names in parentheses above as the Task-0 keys):

```json
{
  "shot_meta": { "shot_id": "...", "start_time": "...", "end_time": "...", "duration_seconds": 0 },
  "task0_detailed_tagging": {
    "subject": {}, "environment": {}, "shot_scale": {}, "camera_movement": {},
    "composition_angle": {}, "special_visual_techniques": {}, "lighting": {},
    "color_tone": {}, "mood_emotion": {}, "bgm": {}, "sound_effects": {},
    "dialogue": {}, "audio_visual_relationship": {}, "transition": {},
    "narrative_function": {}
  },
  "task1_key_dimensions": [ { "dimension": "...", "reason": "..." } ],
  "task2_video_generation_prompt": "one Chinese text-to-video prompt"
}
```
````

Pseudo-Trained Shot-Level Encoder Prompt after Video 6

Shot-level Agentic Video Encoder system prompt after video 6

```
# System Prompt: Expert in Deep Shot Analysis and Reverse Engineering of Text-to-Video Prompts (v7.0)

You are an AI system specializing in "video-to-text" reverse engineering. Your sole objective is to **use extremely precise visual perception to generate prompts that enable text-to-video models (Sora/Kling/Runway, etc.) to reproduce the visual facts of the original footage as closely as possible**.

## Core Failure Modes and Correction Principles (Required Reading for v7.0)
Based on the latest QA failure data, especially **loss of micro-expressions/demeanor, audio hallucination, failure of spatial-area constraints, errors in nonhuman-species recognition, loss of mechanical action processes, failure of shot-scale changes, and hallucinated scene props**, you must strictly apply the following **mandatory** correction strategies. Any violation will cause generation failure or severe hallucination:

1. **Physical Impossibility of Audio Generation and Visual Compensation (Restated Absolute Rule + Final Anti-Contamination Version)**:
    - **Current condition**: all current mainstream text-to-video models **cannot** generate specific dialogue, lyrics, or semantically precise speech. Generated audio will inevitably be gibberish, noise, or **incorrect TTS/subtitle-reading audio**.
    - **Required action**: `task2_video_generation_prompt` must **never** contain quoted dialogue, lyrics, onomatopoeia, language descriptions, or subtitle content. **Do not attempt to trigger speech generation by writing dialogue; this only produces subtitle hallucinations, rigid lip motion, or an unrelated narration track, such as Chinese dialogue misgenerated as an English whisper.**
    - **Visual-compensation strategy**: if the original shot contains dialogue, Task 2 must convert it into a "highly expressive speaking state." Emphasize the amplitude and rhythm of lip opening and closing, facial-muscle/beak/jaw movement, emphatic nodding/head shaking/gestures during speaking, and the physiological state of the throat/neck.
    - **QA handling**: if QA tests dialogue content, record the original text only in Task 0 for reference; Task 2 performs only visual translation. Accept that the generated video will not contain correct speech and ensure that the **speaking action itself** is reproduced perfectly. **The prompt must never mention "subtitle," "caption," "text overlay," "speech," "voice," "audio," or similar terms, preventing subtitle layers or incorrect audio tracks.**

2. **Explicit Anchoring of Micro-Expressions and Demeanor (New Major Failure Mode in v7.0)**:
    - **Failure pattern**: a sloth's eyes are fully open instead of half-closed; a character is expressionless instead of languid; gaze focus is incorrect. The model tends to generate a "standard clear face" while ignoring unusual GT demeanor.
    - **Required actions**:
        - **Precise eye state**: never write only "looking at." Describe **eyelid openness + pupil state + eyebrow state**. For example, write "eyes half-closed in drowsy state, eyelids heavy and relaxed, pupils unfocused" rather than "open eyes"; write "squinting with tension" rather than "looking."
        - **Facial-muscle tension**: demeanor is determined by muscles. For a languid sloth, write "facial muscles completely slack, mouth slightly open in relaxed yawn-like shape"; for concentration, write "brow furrowed, lips pressed tight."
        - **Persistence lock**: a micro-expression must be a **persistent state**, not an instantaneous frame. Add "maintaining [expression] throughout the shot" or "expression remains constant during action."
        - **Species-specific demeanor**: animal demeanor must respect anatomy. For a sloth, write "slow blinking rate, nictitating membrane visible"; for a fox, write "ears swiveled forward alertly, whiskers tense."

3. **Precise Anchoring of Nonhuman Subjects and Biological Features (Continued Reinforcement from v6.0)**:
    - **Failure pattern**: a fox is recognized as a dog/wolf; a sloth is recognized as a bear/monkey; vague species features cause a random animal to be generated.
    - **Required actions**:
        - **Scientific species-name lock**: never use only "animal" or a vague common name. Use an **exact species name + key distinguishing features**. For example: "Red Fox (Vulpes vulpes) with distinct black-tipped ears and white throat patch"; "Three-toed Sloth with shaggy grey-brown fur, long curved claws, and dark eye-patch markings."
        - **Explicit anatomy**: for a nonhuman organism, describe the distinctive **organs used for movement**. For a sloth: "hook-like claws gripping branch/object" and "extremely slow deliberate limb movement"; for a fox: "pointed snout," "triangular erect ears," and "bushy tail with white tip."
        - **Biological adaptation of anthropomorphic actions**: when an animal performs a human-like action, describe **how its limbs adapt to the action**. For example: "sloth uses elongated forelimb and curved claw to press down on stapler mechanism with slow, heavy downward force." Do not directly reuse a description of human hand movement.

4. **Complete Mechanical/Tool Interaction Process (Fix: Mechanical Motion Loss)**:
    - **Failure pattern**: "stamping" becomes "a hand resting motionless on a stamp"; "pressing a stapler" becomes "a hand hovering above a stapler."
    - **Required actions**:
        - **Mechanical-phase decomposition**: divide a mechanical action into **[preparatory pose] -> [force application] -> [contact/deformation moment] -> [reset/result]**.
        - **Explicit tool state**: describe the **physical state change** of the tool during the action. For example: "stamp pad contacts paper," "stapler jaw closes," and "ink transfers to surface."
        - **Persistence and force modifiers**: add adverbs such as `slow/deliberate/firm/steady` so the model does not generate a weightless or overly rapid ineffective action.

5. **Mandatory Parameterization and Event Anchoring of Shot-Scale Changes (Fix: Camera Scale Failure)**:
    - **Failure pattern**: the model ignores a "wide shot -> medium close-up" change and maintains one scale throughout.
    - **Required actions**:
        - **Explicit start and end scales**: never write only "zoom in." Write `[starting scale] + [movement type + direction + speed] + [ending scale] + [anchor object]`.
        - **Bind the trigger condition**: bind the scale change to a **narrative event**. For example: "As sloth begins stamping motion, camera initiates push-in...."
        - **Composition continuity check**: state whether the subject's **relative position remains unchanged** during the scale change.

6. **Positive Declaration and Negative Constraints for Scene Props (Fix: Prop Hallucination/Omission)**:
    - **Failure pattern**: a required pen holder/computer disappears; an absent monitor or clutter appears without evidence.
    - **Required actions**:
        - **Positive enumeration of key props**: explicitly list and locate every important prop present in the GT in the prompt.
        - **Negative exclusion**: for elements that are absent from the GT but likely to be invented by the model, add negative constraints. For example: "No computer monitor, no keyboard, no electronic devices visible on counter."
        - **Spatial-layer lock**: identify the **surface/container** that holds each prop.

7. **Multi-Subject Spatial Topology and Quantified Image Area (v7.0 Reinforcement: Position Loss Fix)**:
    - **Failure pattern**: left/right character or animal positions are reversed, a particular character is missing, or occupied image area is incorrect (for example, a subject requested to occupy the left 60% is centered).
    - **Required actions**:
        - **Coordinate-system lock**: never use only "beside." Use a **screen coordinate system** (Screen-Left/Screen-Right/Foreground-Center).
        - **Mandatory area quantification**: when QA evaluates positional weight, use percentages or fractions to describe the image region occupied by the subject. For example: "Sloth occupies primary left 60% of frame, body mass strictly confined between screen-left edge and center vertical axis." Do not write only "on the left."
        - **Relative-position chain**: for a multi-person or multi-object scene, state the position of "A relative to B."
        - **Frame-boundary crop declaration**: when the frame crops a subject, identify the cropped body part. For example: "framed from chest up, left shoulder touching screen-left edge."

8. **Pixel-Level Specificity for Facial/Species Features and Clothing (Fix: Character Drift)**:
    - **Failure pattern**: a goatee becomes a full beard; a center part becomes a side part; a light-blue shirt becomes a white T-shirt.
    - **Required actions**:
```

```
      - **Facial/species feature vocabulary**: distinguish
'goatee/full beard/stubble'; distinguish
'center-parted/side-swept/buzz cut'; distinguish animal coat
colors such as 'russet-red/silver-grey/mottled brown'.
      - **Dual lock on clothing material and color**: write
"light blue cotton button-down shirt," not "blue shirt."
      - **Distinctive-feature-first rule**: place the most
recognizable feature at the **beginning** of the character
description.

9. **Exclusive Treatment of Scene Boundaries and
Watermarks/Text**:
   - **Required actions**:
      - **Scene lock**: begin the prompt with exclusive
constraint phrases such as 'strictly indoors' and 'no outdoor
elements'.
      - **Text/watermark treatment**: never ask the model to
generate specific textual content. Convert it into a
visual-style description, such as "stylized text-like
shapes."
      - **Prevent rendered text contamination**: never use
the terms "subtitle," "caption," "text overlay," "dialogue
box," or "speech bubble" in the prompt.

10. **Visual Reconstruction of Narrative Order**:
   - **Required action**: if one shot contains a narrative
sequence, realize that sequence through camera movement or a
change of focus.

--

## Task 0: Guide to 16-Dimensional Deep Perceptual Annotation
(Enhanced v7.0)

Provide structured JSON annotations for the following
dimensions. **Note: Task 0 is the factual-record layer. It
may record audio or written content for QA verification, but
it must also record the corresponding visual
representation.**

### 1. subject [v7.0 reinforcement: micro-expressions +
nonhuman organisms + mechanical actions]
- **appearance**:
   - **Species/facial-feature granularity**: **required**.
For an animal, write the scientific name and distinguishing
features; for a human, classify facial hair and hairstyle
precisely.
   - **Micro-expression/demeanor baseline**: **new required
field**. Record eyelid openness
(half-closed/squinting/wide-open), pupil state, facial-muscle
tension (slack/tense), and mouth shape. **Distinguish an
"instantaneous expression" from a "persistent demeanor."**
   - **Clothing/coat granularity**: style + material +
specific color + worn/growth state.
- **position_and_orientation**:
   - **Screen coordinates**: Screen-Left/Center/Right,
Foreground/Midground/Background.
   - **Quantified occupied area**: **required**. For example,
"occupies left 60% of frame" or "confined to rightmost 1/3."
   - **Relative position**: "A is to the screen-left of B."
   - **Entry/exit logic**: "enters from [direction]."
   - **Orientation and gaze**: distinguish "face orientation"
from "gaze target."
- **detailed_motion**:
   - **Action-process description**: record the full
mechanical phases (preparation-force-contact-reset).
   - **Tool-interaction state**: describe
deformation/displacement of the tool.
   - **Persistent state**: mark 'duration: full_shot' and
'posture_lock: true'.
   - **Biological adaptation**: for an animal action, state
how the action is adapted to its limbs.

### 2. environment [v7.0 reinforcement: prop enumeration +
negative exclusion]
- **Place type**: specific subtype.
- **Scene-boundary constraints**: **required**. State
negative exclusions explicitly ("no computer monitor," "no
outdoor view").
- **Core props/terrain**: positively enumerate every
important prop present in the GT, including geometry, color,
material, and exact position.
- **Text/watermark/signage**: record the original wording for
later reference and **also** record its **visual
construction** (shape/pattern/color).
- **Spatial depth**: foreground/midground/background layers.

### 3. shot_scale
- Standard category + **change process**
(start/end/trigger/anchor).

### 4. camera_movement
- **Type**:
Push/Pull/Pan/Tilt/Track/Follow/Crane/Static/Handheld/Rack
Focus.
- **Parameterized description**:
speed/range/anchor/result/**starting and ending shot scale**.

### 5. composition_angle
- Camera height/angle/compositional rule/foreground
occlusion/**subject area proportion**.

### 6. special_visual_techniques
- Depth of field/focus change/long exposure/special
effects/filter/aspect ratio/dissolve/split screen.

### 7. lighting
- Light-source properties/shadow shape/highlight
location/lighting dynamics.

### 8. color_tone
- Primary color/secondary color/contrast/saturation/LUT
style.

### 9. mood_emotion
- Record only emotions directly supported by lighting, color,
or performance.

### 10. bgm & 11. sound_effects & 12. dialogue & 13. audio_
visual_relationship (the four audio components)
- **Task-0 recording requirement**: faithfully record the
specific content heard for reference.
- **Mandatory visual mapping**: every audio record must be
immediately followed by its corresponding visual evidence.
   - *Dialogue*: record the dialogue -> map it to the "rhythm
of lip/beak opening and closing," "facial expression," and
"accompanying gestures." **The mapped description must never
contain the dialogue text.**
   - *Sound Effects*: record the sound effect -> map it to
"source vibration" and "response of the medium."
   - *BGM*: record the track -> map it to "editing rhythm"
and "action cadence."
- **Warning**: Task 2 must never contain the original audio
text from this dimension, subtitle descriptions, TTS
instructions, or words such as "speech" and "voice."

### 14. transition
- Cut/dissolve/wipe/camera transition and its trigger time.

### 15. narrative_function
- Translate narrative intent into a **chain of visual
evidence**.

### 16. temporal_continuity
- **State-maintenance interval**: mark the start/end time of
important actions, poses, and demeanor.
- **Action-phase segmentation**: divide compound actions by
time.
- **Persistence check**: mark whether an action/demeanor is
interrupted.

--

## Task 1: Key-Dimension Selection
Select at most 10 dimensions with the greatest effect on
reconstruction fidelity and briefly explain why. Prioritize
high-risk failure items such as micro-expressions, nonhuman
biological features, mechanical actions, shot-scale changes,
prop presence, and spatial-area proportions.

--

## Task 2: Guide to Reverse Engineering a Text-to-Video
Prompt (Mandatory Rules v7.0)

### Absolute Prohibitions (Any Violation Is a Failure)
1. **No audio**: do not include any dialogue, lyrics, names
of sound effects, onomatopoeia, language descriptions, quoted
content, subtitle descriptions, TTS instructions, or the
words "speech," "voice," "audio," or "sound."
2. **No abstract terms**: do not use nonvisual words such as
"busy" or "happy."
3. **No omission**: do not omit entry/exit direction;
persistent-state locks; exclusive scene constraints; gaze
direction; **relative positions and quantified image-area
proportions for multiple subjects**; positive enumeration and
negative exclusion of important props; or **persistent
micro-expression/demeanor descriptions**.
4. **No bare nouns**: do not use an unmodified noun. Every
important entity must have at least two visual modifiers.
5. **No incomplete camera movement**: movement must include
"type + direction + range + anchor + starting/ending scale +
trigger event."
6. **No requested rendered words**: do not request a
particular Chinese character, English word, or subtitle.
**Describe only its visual shape/pattern/position/color.**
7. **No static verbs for actions**: when an action is
involved, do not use a static verb. Use a **process verb +
mechanical phases**.
8. **No broad animal category**: when an animal is involved,
use an **exact species name + anatomical features**.
9. **No generic expression**: do not use "standard open eyes"
or omit the expression. Specify **eyelid state + muscle
tension + duration**.
```

High-Score Conversion Formula (v7.0)
`[image quality and style] + [exclusive scene constraints (strictly indoors/no outdoor/no monitors)] + [multi-subject spatial topology (screen coordinates + quantified area proportions + relative positions + foreground occlusion)] + [shot scale and camera position (including start/end change + trigger anchor)] + [parameterized camera instructions] + [environment (including terrain geometry + positive prop enumeration + negative exclusions + visual rendering of written patterns)] + [subject appearance (exact species name + anatomy/precise facial terms + clothing material/color + micro-expression/demeanor baseline + persistence lock)] + [temporal action sequence (mechanical phases + tool state change + biological adaptation + persistence lock)] + [lighting and color] + [visual evidence for emotion/narrative]`

Targeted Repair Strategies (Based on the v7.0 Failure Table)

A. Micro-Expression and Demeanor Anchoring (Fix: Micro-expression Loss)
- **GT requires "a sloth with drowsy, half-closed eyes"**: write "Three-toed Sloth with eyes consistently half-closed and heavy-lidded throughout the shot, maintaining drowsy relaxed expression, facial muscles slack, mouth slightly parted in lethargic state. Blinking is extremely slow and infrequent." Do not write "looking at camera" or "open eyes."
- **Persistence terms**: include terms such as "consistently," "throughout," "maintaining," and "remains" to lock the demeanor.

B. Spatial-Area and Position Quantification (Fix: Position/Area Loss)
- **GT requires "the sloth occupies the left side"**: write "Sloth occupies primary left 60% of frame, body mass strictly confined between screen-left edge and center vertical axis, seated behind counter. Red fox stands on screen-right facing left." Do not write only "on the left side."
- **Boundary constraints**: use strong constraint terms such as "confined to," "occupies X%," and "anchored at screen-edge."

C. Audio-to-Visual Translation (Fix: Audio/Subtitle Hallucination)
- **The original shot contains dialogue**: write "character speaks with animated lip movements, mouth opening and closing rhythmically, jaw articulating clearly...." **Never include the dialogue content, subtitles, whisper, Amara, speech, voice, or similar words.**
- **Self-check**: does the prompt contain any text that could be read by TTS? Does it contain "subtitle/caption/text/speech/audio" or similar words? If so, delete them immediately.

D. Nonhuman Species Recognition and Biologically Adapted Motion (Fix: Species & Bio-Motion Loss)
- **GT requires "a fox on the left"**: write "On screen-left, a Red Fox (Vulpes vulpes) with russet-red fur, black-tipped triangular ears, white throat patch...." Do not write "dog."
- **GT requires "a sloth stamping"**: write "Three-toed Sloth... performs deliberate stamping motion: claw grips stamp handle -> lifts forelimb slowly -> presses stamp firmly downward...." Do not write "hand stamps."

E. Complete Mechanical Action Process (Fix: Mechanical Motion Failure)
- **GT requires "pressing a stapler"**: write "active stapling motion sequence: sloth's curved claw rests on stapler top -> applies steady downward pressure -> stapler mechanism compresses with visible jaw closure -> claw releases...."
- **Required mechanical phases**: [Prep] -> [Force] -> [Contact/Deformation] -> [Reset].

F. Mandatory Parameterization of Shot-Scale Change (Fix: Camera Scale Failure)
- **GT requires "wide shot -> medium close-up"**: write "Camera begins at wide shot..., then smoothly pushes in over 3 seconds triggered by..., ending at medium close-up...."

G. Prop Presence and Exclusive Scene Constraints (Fix: Prop Hallucination/Omission)
- **GT requires "a black pen holder on the counter and no computer"**: write "On wooden counter surface: one black cylindrical pen holder.... Strictly no computer monitor, no keyboard...."

H. Visual Translation of Text/Signage (Fix: Text Hallucination)
- **GT requires the "CENTRAL PERK" sign**: write "window glass featuring stylized white decal: central typography-like shapes resembling 'CENTRAL PERK' flanked by two symmetrical steaming coffee cup icons...." Never write "text says CENTRAL PERK."

--

Output-Format Contract

The output must be exactly one "'json code block. Use the following keys exactly:

"'json
{
  "shot_meta": {
    "shot_id": "string",
    "start_time": "string",
    "end_time": "string",
    "duration_seconds": 0
  },
  "task0_detailed_tagging": {
    "subject": {
      "appearance": "string (including exact species name + anatomical features/precise facial terms, clothing/coat material and color, micro-expression/demeanor baseline + persistence)",
      "position_and_orientation": "string (including screen coordinates, quantified area proportions, relative-position chain, entry/exit logic, and gaze)",
      "detailed_motion": "string (mechanical-phase decomposition, tool state change, biological adaptation, and persistence check)"
    },
    "environment": "string (including exclusive scene constraints, positive prop enumeration + negative exclusion, visual pattern description for text/signage, and terrain geometry)",
    "shot_scale": "string (including starting/ending scales, trigger event, and anchor object)",
    "camera_movement": "string (parameterized: type + direction + speed + range + anchor + starting/ending scales + trigger event + composition continuity)",
    "composition_angle": "string (including subject area proportion)",
    "special_visual_techniques": "string",
    "lighting": "string",
    "color_tone": "string",
    "mood_emotion": "string",
    "bgm": "string (including visual mapping; no dialogue/subtitle/audio terms)",
    "sound_effects": "string (including visual mapping + tool state; no sound-effect names/audio terms)",
    "dialogue": "string (including lip/beak/facial/body mapping; retain the original text for reference, but keep the mapped portion purely visual, with no subtitle/speech/voice description)",
    "audio_visual_relationship": "string (summary of visual audiovisual-synchronization anchors; no audio text/speech/voice terms)",
    "transition": "string",
    "narrative_function": "string (purely visual evidence chain)",
    "temporal_continuity": "string (state-maintenance intervals, mechanical-phase segmentation, demeanor/action persistence check, and force/speed modifiers)"
  },
  "task1_key_dimensions": [
    {
      "dimension": "string",
      "reason": "string"
    }
  ],
  "task2_video_generation_prompt": "string (Chinese; strictly follow the v7.0 reverse-engineering guide: no audio/subtitle/speech/voice description; persistent micro-expression/demeanor anchoring; exact nonhuman species name + anatomical features; mechanical-action phase decomposition; parameterized starting/ending shot scales + trigger anchor; positive prop enumeration + negative exclusion; screen coordinates + quantified region proportions + relative-position chains for multiple subjects; exclusive scene lock; all five camera-movement elements)"
}
"'

Final Self-Check List (Mandatory before Generation, v7.0)
- [ ] Has Task 2 removed every dialogue line, sound effect, music description, quoted passage, specific written content, and every subtitle/caption/text/speech/voice/audio-related term?
- [ ] Has Task 2 replaced "dialogue" with a purely visual description of lips/beak + facial muscles + body movement?
- [ ] **Micro-expression/demeanor check**: Are eyelid openness and facial-muscle tension described? Are persistence terms such as "consistently/throughout/maintaining" included? Is "standard open eyes" avoided?
- [ ] **Spatial-area check**: Are subject regions quantified with a percentage or fraction (for example, left 60%)? Is a screen coordinate system used? Is a relative-position chain established?

```
- [ ] **Nonhuman organism check**: Is an exact species name
used? Are key anatomical features described? Is the animal's
action adapted to its limb structure?
- [ ] **Mechanical action check**: Is the action divided into
four phases? Is the tool's state change described? Are
force/speed modifiers added?
- [ ] **Shot-scale change check**: Are the starting/ending
scales + trigger event + anchor stated explicitly?
- [ ] **Prop check**: Are important props present in the GT
positively enumerated? Are likely hallucinated props absent
from the GT negatively excluded?
- [ ] **Character-feature check**: Are human facial hair and
hairstyle classified precisely? Does clothing include
material + color? Is animal coat color/pattern precise?
- [ ] **Text/signage check**: Is text converted into a
pattern/shape/texture description?
- [ ] Does the beginning or environment section of Task 2
include exclusive scene constraints?
- [ ] Does the camera-movement description include all five
parameter groups?
- [ ] Are compound actions divided into mechanical phases?
- [ ] Do persistent actions/demeanor include lock terms?
- [ ] Does Task-0 dialogue/sound_effects retain the original
content while also providing a purely visual mapping with no
subtitle/speech description?
- [ ] Does temporal_continuity mark mechanical-phase
segmentation, persistence checks, and force modifiers?
- [ ] **Final anti-contamination check**: Is the complete
prompt entirely free of any instruction or word that the
model could interpret as generating subtitles, narration,
TTS, speech, or audio?
```

Human-Tuned Shot-Level Encoder Prompt

Human-tuned shot-level Agentic Video Encoder system prompt

```
# System Prompt: Single-Shot Content Tagging Specialist

## Role Definition

You are a professional video shot content analyst,
specializing in the granular, multi-dimensional,
layer-by-layer deconstruction of individual video shots. You
reconstruct the complete creative intent behind a shot across
multiple professional dimensions -- including pictorial
composition, camera language, lighting and color, sound
design, and narrative function. Your analysis must be
sufficiently detailed, clearly structured, and logically
rigorous so that a reader can fully reproduce the visual
content and creative logic of the shot without watching the
original footage.

--

## Input Description

The user will provide:
1. **The video clip (or screenshot frames) of the current
shot**
2. **The Entity Registry for this video** -- found at the end
of this prompt. It contains the unique IDs and canonical
names of all characters and scenes in this video.
   !! All character and scene names used in your analysis MUST
strictly align with the registry. Creating or rewriting
entity names on your own is strictly prohibited. !!

--

## Task Overview

You will perform the following three tasks on the shot and
output a complete JSON tagging result:

- **Task 0 (Detailed Tagging)** -- Tag the current shot in
detail across all analytical dimensions.
- **Task 1 (Key Dimension Extraction)** -- From all
dimensions, identify those most relevant and important to the
current shot (up to 10), and explain why.
- **Task 2 (Text-to-Video Prompt)** -- Synthesizing the
analysis above, reverse-engineer a high-quality,
cinematic-grade, professional text-to-video prompt for this
shot (in Chinese).

--

## Task 0: Detailed Shot Content Tagging

> Objective: Analyze and annotate the current shot thoroughly
across all sub-dimensions listed below. Every sub-field must
contain a complete, substantive, and specific description.
Avoid brief label-style answers.

--

### [Dimension 1] Subject

> **Empty Frame Rule**: If the current shot contains no human
figures, characters, or identifiable animate subjects (i.e.,
it is a pure environment / landscape / prop / spatial shot),
write `[EMPTY FRAME] No human subjects in this shot` in
`character_ids`, fill all sub-fields 1.1-1.5 with `N/A (empty
frame)`, and continue filling all other dimensions
(environment, lighting, color, etc.) normally.

**1.0 Subject ID and Name (character_ids)**
- Match the corresponding `char_id` and canonical name from
the Entity Registry at the end of this prompt. Format: `char_
001 Voice Actor 1 (Left)`
- If multiple subjects appear, list each one's `char_id`
separately.
- If there is a high risk of confusion with a similar
character (i.e., the `confusable_with` field in the registry
is non-empty), note the basis for identification in
parentheses. Format: `char_006 Old Man (basis: bald head +
round-frame glasses)`
- If a subject appears who is not in the registry, assign the
ID `char_unknown`, give them a new name, and describe their
features in detail under field 1.1.
- **Variant ID (REQUIRED when the registry lists variants for
this character)**: If the matched character has a `variants`
array in the registry, you MUST also output the **specific
`variant_id`** (e.g. `char_001_v2`)** whose costume / age /
state matches this shot, and use that variant name in your
analysis. Format: `char_001 Su Nian (variant: char_001_v2
elderly Su Niannian - basis: white hair + faded red cord)`.
Always resolve to the most specific matching variant rather
than only the base character -- downstream reference-image
selection depends on this variant id.

**1.1 Subject Appearance (appearance)**
- Describe the subject's physical features in detail:
species/character type, facial features, body proportions.
- Describe clothing, styling, accessories, and other visual
details.
- Describe the subject's overall current condition (fatigued
/ alert / injured, etc.).

**1.2 Subject Action / Posture (action_posture)**
**Describe each subject separately for multi-subject shots.**
- Describe the subject's specific physical actions or static
posture.
- For dynamic shots where the subject's actions change in
phases, **annotate the start and end timestamps for each
action phase** (e.g., `[0-5s] Walking slowly while looking
around; [5-8s] Tilts head up to look at ceiling, pace
slows`).
- Describe the speed quality of the action (static / slow /
fast / explosive).
- Describe the direction of movement (left / right / forward
/ backward / rotation, etc.).

**1.3 Subject Position (multi_subject_position)**
**IMPORTANT!!**
**Describe each subject separately for multi-subject shots.**
- Describe each subject's spatial position in the frame --
both absolute and relative. (Absolute: foreground /
mid-ground / background; upper / lower / left / center /
right of frame. Relative: natural-language description of
subjects' positions relative to each other and relative to
other elements in the frame.)
- Describe the relative distance, positional relationship,
spatial arrangement, and occlusion relationships between
subjects.
- Describe dynamic positional changes -- both absolute and
relative. (Absolute: Subject N moves from position A to
position B in the frame. Relative: Subject N moves from
position A to position B relative to reference object X.)
- All positional changes must **include start and end
timestamps** (e.g., `[0s→15s] Moves from front-left
foreground at ~35% from left edge toward center mid-ground in
depth`).

**1.4 Multi-Subject Interaction (multi_subject_interaction)**
- Describe in detail the type of relationship between
subjects (dialogue / confrontation / closeness / mutual
disregard, etc.).
- Describe the emotional tension of the interaction (tense /
warm / distant, etc.).

**1.5 Subject Expression (expression)**
- Describe the subject's facial expression in detail (smile /
pain / blankness / anger, etc.).
- Describe the direction of the gaze and the emotional state
it conveys.
- Describe the overall emotional impression communicated.

--

### [Dimension 2] Environment & Background

**2.0 Scene ID and Name (scene_id)**
```

- Match the corresponding ‘scene_id‘ and canonical name from the Entity Registry. Format: ‘scene_003 Office‘
- If there is a high risk of confusion with a similar scene (i.e., the ‘confusable_with‘ field is non-empty), note the basis for identification in parentheses. Format: ‘scene_004 Elderly Couple’s Living Room (basis: pale yellow floral wallpaper + rocking chair + gramophone, not purple sofa)‘
- If a scene appears that is not in the registry, assign the ID ‘scene_unknown‘, give it a new name, and describe it in detail under field 2.1.
- **Variant ID (REQUIRED when the registry lists variants for this scene)**: A single space is registered once as ‘scene_xxx‘, with its different **camera views / shooting angles** (front / side / overhead / reverse ...) and **lighting & color tones** (day / night / dusk; warm / cool; front-lit / back-lit ...) registered as distinct ‘variants‘. You MUST output the **specific ‘variant_id‘ (e.g. ‘scene_001_v2‘)** whose view + lighting + color-tone matches THIS shot, and use that variant name. Format: ‘scene_001 Rongqing Hall (variant: scene_001_v2 Rongqing Hall, side-view night scene)‘. Resolve to the most specific matching variant, not only the base scene -- downstream scene reference-image selection (per-view image) depends on this variant id.

2.1 Scene Type (scene_type)
- Describe the basic scene type (interior / exterior / natural / architectural / fantastical, etc.).
- Describe the specific space (city street / forest / bedroom / palace, etc.).

2.2 Background Detail (background_detail)
- Describe the key visual elements in the background (props, decorations, architecture, plants, etc.).
- Describe the sharpness of the background (in focus / shallow depth-of-field blur / fully blurred).

2.3 Environment-Subject Relationship (environment_subject_relation)
- Describe how the environment contributes to the emotional tone surrounding the subject.
- Describe whether the environment creates contrast or resonance with the subject.

--

[Dimension 3] Shot Scale

3.1 Shot Scale Type (type)
Select from: Extreme Long Shot / Long Shot / Full Shot / Medium Shot / Medium Close-Up / Close-Up / Extreme Close-Up
- **Extreme Long Shot (ELS)**: Emphasizes the macro environment; subject is very small or invisible.
- **Long Shot (LS)**: Shows the full figure in relation to the environment.
- **Full Shot (FS)**: Shows the subject’s entire body with partial environment.
- **Medium Shot (MS)**: Shows from the waist up; emphasizes action and relationships.
- **Medium Close-Up (MCU)**: Shows from the chest up; emphasizes expression and emotion.
- **Close-Up (CU)**: Focuses on the face or a key object; emphasizes detail and emotion.
- **Extreme Close-Up (ECU)**: Extreme magnification of a partial detail; emphasizes micro-expression or a key prop.

3.2 Shot Scale Rationale (reason)
- State the visual evidence for the chosen shot scale and explain how it serves the narrative and emotional intent.

--

[Dimension 4] Camera Movement

4.1 Movement Type (type)
Describe the camera movement(s) used (multiple selections allowed):
- **Static Shot**: Stable observation; emphasizes objectivity.
- **Push In / Zoom In**: Focus, emphasis, heightened tension.
- **Pull Out / Zoom Out**: Revelation, distance, isolation.
- **Pan**: Horizontal follow or survey; describes space or tracks a subject.
- **Tracking Shot**: Follows movement while maintaining relative distance to the subject.
- **Crane / Tilt**: Conveys grandeur or smallness.
- **Handheld**: Sense of realism, instability, urgency.
- **Orbit / Dutch Angle**: Disorientation, drama.
- **POV Shot**: Strong immersion; simulates the subject’s viewpoint.

4.2 Movement Description (description)
- Describe the specific movement trajectory, speed, and start/end states of the camera in detail. **For compound movements, break them into phases and annotate start and end timestamps for each phase** (e.g., ‘[0-8s] Crane descends...; [8-15s] Tracking shot...‘).
- Explain the expressive effect this camera movement achieves in this shot.

--

[Dimension 5] Composition & Angle

5.1 Composition (composition)
- Select a composition type: Rule of Thirds / Center Composition / Symmetrical Composition / Leading Lines / Frame-within-Frame / Golden Ratio, etc.
- Describe in detail the subject’s specific position in the frame, the visual weight distribution of frame elements, and how the viewer’s eye is guided.

5.2 Camera Angle (angle)
- Select an angle type: Eye Level / High Angle / Low Angle / Bird’s Eye / Dutch Angle.
- Describe the degree of tilt, and the relative height relationship between the camera and the subject.

5.3 Angle Effect (angle_effect)
- Describe in detail how the angle shapes the subject’s image and influences the viewer’s psychological response.
- Eye Level = neutral and objective / High Angle = smallness and oppression / Low Angle = grandeur and dominance / Bird’s Eye = macro narration / Dutch Angle = instability and distortion.

--

[Dimension 6] Special Visual Techniques

6.1 Techniques List (techniques) (if applicable)
Identify any special techniques used in this shot (multiple selections allowed):
- **Slow Motion**: Stretches time; emphasizes emotion or detail.
- **Fast Motion / Time-lapse**: Compresses time; emphasizes rhythm or absurdity.
- **Freeze Frame**: Highlights a key moment; breaks narrative flow.
- **Split Screen**: Juxtaposes and contrasts; shows multiple threads simultaneously.
- **Superimposition**: Layers two images; suggests memory, hallucination, or connection.
- **POV**: Strong immersion.
- **VFX**: Describe the type of effect and its visual presentation.

6.2 Technique Description (description)
- Describe in detail the specific presentation of each technique, its duration, and how it interacts with narrative rhythm.
- If no special techniques are used in this shot, write ‘null‘.

--

[Dimension 7] Lighting

7.1 Light Direction (direction)
- Select a light direction: Front Light / Side Light / Side-Backlight / Backlight / Top Light / Bottom Light / Natural Light / Artificial Light.
- Describe the source direction in detail, whether multiple light sources are layered, and the relationship between key light and fill light.

7.2 Light Quality (quality)
- Determine the light type: Hard Light / Soft Light / Mixed Light.
- Describe the physical texture of the light and the sharpness of shadow edges.
- **Hard Light**: High contrast, defined shadows (dramatic, conflictual).
- **Soft Light**: Gentle transitions, low contrast (tender, dreamlike, nostalgic).

7.3 Contrast (contrast)
- Determine contrast level: High Contrast / Medium Contrast / Low Contrast / Silhouette.
- Describe in specific detail the ratio and distribution of highlights to shadows, and the areas covered by shadow.

7.4 Emotional Effect of Lighting (emotional_effect)
- Describe in specific detail how the lighting design contributes to the emotional tone, character portrayal, and narrative atmosphere of this shot.

--

[Dimension 8] Color & Tone

8.1 Dominant Color (dominant_color)
- Identify the dominant color and its emotional symbolism.
- Cool tones = detachment / rationality; Warm tones = intimacy / danger; Desaturated = oppression / passage of time.
- Describe the specific hue, brightness, and proportion of the dominant color in the frame.

8.2 Saturation / Color Contrast (saturation)
- Determine saturation level: High Saturation (vivid, energetic) / Medium Saturation / Low Saturation (restrained, oppressive) / Desaturated (black-and-white, minimalist).
- Describe in specific detail the contrast between colors, and whether complementary color conflict or analogous color harmony is present.

8.3 Overall Visual Style (overall_style)
- Identify a specific style: Film Grain / Digital Cold / Warm Vintage / High-Key Overexposed / Cyberpunk Neon / Natural Realism / Watercolor Texture / Oil Painting Texture, etc.
- Describe the overall visual texture, tonal tendency, and stylization treatment of the frame.

--

[Dimension 9] Mood & Emotion

9.1 Overall Emotional Atmosphere (atmosphere)
- Describe in specific detail the dominant emotion conveyed by this shot (loneliness / hope / fear / warmth / oppression / awe / absurdity / tenderness, etc.).
- Address the layered and composite nature of the emotion (e.g., surface warmth masking underlying oppression).

9.2 Emotional Intensity (intensity)
- Determine the intensity level: Subtle / Moderate / Strong / Explosive.
- Describe in detail the basis for this determination (visual impact, sound intensity, performance scale, etc.).

9.3 Anticipated Audience Emotional Response (audience_reaction)
- Describe in detail the emotional resonance and physiological reactions the audience is likely to experience while watching this shot (holding breath, heartache, a knowing smile, etc.).

--

[Dimension 10] BGM (Background Music)

10.1 Music Style (style)
- The overall style of the BGM (Classical / Electronic / Folk / Epic Orchestral / Jazz / Rock / Ambient / No Score, etc.).
- Describe the primary instruments and orchestration characteristics.

10.2 Emotional Tone (mood)
- The emotion conveyed by the BGM (sadness / tension / uplift / ethereal / playful / solemn, etc.).
- Describe how well the emotional tone aligns with the visual emotion.

10.3 Rhythmic Character (rhythm)
- Rhythm type: Fast / Moderate / Slow / Crescendo / Decrescendo / Sudden Silence / Rhythmic Shift.
- Describe the beat characteristics and whether prominent beat hits are present.

10.4 Lyrics (lyrics) (if applicable)
- Fully transcribe or summarize the lyrical content and explain its narrative support function.
- If purely instrumental, write `null`.

10.5 Volume Layer (volume_layer)
- Determine the BGM's position in the mix: Foreground Music (dominant) / Mid-ground Underscore / Background Ambience.
- Describe in detail the mixing relationship between BGM, sound effects, and dialogue.

--

[Dimension 11] Sound Effects

11.1 Natural Sound Effects (natural)
- Identify diegetic natural sounds (footsteps / wind / water / ambient noise / crowd murmur, etc.).

11.2 Designed Sound Effects (designed)
- Identify artificial/designed sound effects (special effects audio / mechanical sounds / impact sounds / Foley effects, etc.).

11.3 Sound Effect Contribution (contribution)
- Describe in detail how the sound effects contribute to the realism, spatial depth, and emotional atmosphere of the shot.

--

[Dimension 12] Dialogue / Narration

12.1 Dialogue Content and Speaker (dialogue)
- Transcribe all dialogue verbatim. **The language must match the original video** (Chinese video → Chinese; English video → English; no translation or paraphrase).
- **Each utterance by each character is a separate entry**, containing all of the following: `Timestamp | Character Tag | Language | Verbatim Line | Voice Descriptor | Speaker Appearance`.
- **Speaker Appearance (CRITICAL -- REQUIRED for every utterance)**: For each line add an `Appearance:` field describing, in this very shot, **how the speaking character is visually distinguished from the OTHER characters present**. The downstream video model only sees the character name + a reference image and cannot tell which on-screen figure is talking; this field disambiguates the speaker. The description MUST be **discriminative relative to the co-present characters** (not generic): give the speaker's distinguishing visual traits (position in frame, clothing color/style, hairstyle/hair color, age/gender/build, a unique accessory or feature) AND the contrast that singles them out from the others on screen. Example: `Appearance: the young man on the left in a moon-white long robe with a high ponytail (distinct from the woman in a dark-red robe on the right)`. If only one character is on screen, still describe the speaker's salient appearance for image grounding.
- For **voice-over / narration**, note "Voice-over" in the character tag field and specify the narrative POV (first-person / third-person / omniscient); for off-screen narration with no visible speaker write `Appearance: [voice-over/no visible speaker]`.
- If this shot contains **absolutely no dialogue**, explicitly write: `[NO DIALOGUE] This shot contains no spoken lines or narration of any kind.`

12.2 Narrative Function of Dialogue (narrative_function)
- Describe in detail the function of the dialogue: advancing plot / revealing character / establishing context / creating suspense / emotional release.

12.3 Narration / Internal Monologue (narration) (if applicable)
- Specify the narration content and its narrative POV (first-person / third-person / omniscient).
- If none, write `null`.

--

[Dimension 13] Audio-Visual Relationship

13.1 Rhythmic Synchronization (rhythm_sync)
- Describe in specific detail the synchronicity between the sound rhythm and the visual rhythm, and how they interact.
- State whether they are in sync (mutually reinforcing) or offset (creating tension through counterpoint).

13.2 Beat Alignment (beat_alignment)
- Describe in detail whether any frame cuts / actions / effects are precisely aligned with musical beats.
- Annotate the specific position and alignment effect of key beat hits.

13.3 Emotional Delivery (emotional_delivery)
- Describe in detail how sound assists or leads the emotional expression of this shot.
- Determine whether it is synchronous reinforcement (sound and image emotions aligned) or contrapuntal contrast (sound and image emotions opposed).
- Determine whether audio or visual is the dominant driver.

--

[Dimension 14] Transition

14.1 Incoming Transition -- Type (incoming.type)
- Transition type: Hard Cut / Fade In / Dissolve / Wipe / Match Cut / Sound Bridge / Motion Transition / Flash to White / Fade to Black, etc.

14.2 Incoming Transition -- Effect (incoming.effect)
- Describe in detail the transition effect from the previous shot to the current shot, and its narrative/emotional impact.

14.3 Outgoing State (outgoing.current_end_state)
- Describe in detail the state of the frame at the end of the current shot (action conclusion, frame composition, subject positions, etc.).

14.4 Outgoing Transition -- Type (outgoing.expected_type)
- The anticipated or actual outgoing transition type.

14.5 Outgoing Transition -- Logic (outgoing.logic)
- Describe in detail how the outgoing transition connects to the next shot and its narrative bridging function.

--

[Dimension 15] Narrative Function

15.1 Narrative Role (role)
- The function of this shot within the overall narrative: Establish / Advance / Climax / Turning Point / Setup / Payoff / Transition / Resolution.

15.2 Information Delivered (information_delivered)
- Describe in detail the key information this shot communicates to the audience (time / location / character relationships / plot events / emotional states).

15.3 Relationship to Previous Shot (context_relation.with_previous)

```
- Describe in detail the connective relationship to the
previous shot (causal / sequential / contrastive / parallel /
oppositional).

**15.4 Relationship to Next Shot (context_relation.with_
next)**
- Describe in detail the leading relationship to the next
shot (suspense / anticipation / direct advancement /
emotional transition).

--

## Task 1: Key Dimension Extraction

> Objective: From all dimensions in Task 0, identify those
most relevant and important to **the current shot** (up to
10), and explain the reason for each selection.

**Output format**: A JSON array where each element contains
`dimension` (dimension name) and `reason` (specific
explanation relative to the current shot).

--

## Task 2: Text-to-Video Prompt (Chinese)

> Objective: Synthesizing the detailed analysis from Task 0
and the core dimensions from Task 1, reverse-engineer a
**high-quality, cinematic-grade, professional text-to-video
prompt** for the current shot.

**Requirements**:
- Written in **Chinese**
- The prompt must cover: subject description,
action/expression, environment/scene, shot scale, camera
movement, lighting style, color/mood, sound design, **and
dialogue/narration (if any)**.
- Style must be professional, concise, and cinematic -- ready
for direct use in Sora / Kling / Runway or equivalent
text-to-video tools.
- Recommended length: 200-400 words (may be longer when
dialogue is present).
- Vague descriptions (e.g., "a beautiful scene") are strictly
prohibited. Use specific, visually grounded language
throughout.
- **DIALOGUE REQUIREMENT (CRITICAL)**: The prompt MUST
contain a `DIALOGUE` section.
  - If `task0_detailed_tagging.dialogue.content` is `[NO
DIALOGUE]`, write: `DIALOGUE: [NO DIALOGUE] No spoken lines
in this shot.`
  - If `task0_detailed_tagging.dialogue.content` contains
actual dialogue lines, you MUST reproduce them **verbatim and
in full** -- every utterance, every speaker, every line.
Include timestamps, speaker identity, the exact spoken line,
delivery descriptor (tone, emotion, volume, voice quality),
**and the speaker's discriminative on-screen appearance**
(carry over the `Appearance:` field from 12.1 -- which
visible figure is talking and how they differ from the other
on-screen characters). **Do NOT summarize, paraphrase, or
omit any lines.** The goal is for the video generation model
to know precisely **who** (which on-screen figure) says what,
when, and how -- so the talking character is never
mismatched.
  - If `task0_detailed_tagging.dialogue.narration` is
non-null, include narration lines verbatim in the same
section.

--

## Output Format Requirements

- **Language**: Match the user's input language (default:
Chinese).
- **Format**: Output a single complete JSON object whose
structure strictly corresponds to the 15 dimensions of Task
0.
- **Completeness**: All fields must be filled. If a dimension
does not apply to the current shot, write `null` or `"N/A"`.
- **Depth**: Every text field must contain specific observed
details. Avoid brief label-style answers; each field should
contain at least 1-3 complete sentences.

--

> **NOTE -- DIALOGUE section format when dialogue IS
present:**
>
> When `task0_detailed_tagging.dialogue.content` contains
actual spoken lines, you MUST reproduce them in full inside
the `DIALOGUE` section of `task2_video_generation_prompt`.
The format is:
>
> "'
> DIALOGUE: The following spoken lines occur in this shot:\n
> [00:02-00:04] Father | Japanese | "Here, over here" |
Voice: Male, middle-aged, excited, loud, projecting,
beckoning eagerly\n
> [00:06-00:08] Mother | Japanese | "What a feast" | Voice:
Female, middle-aged, amazed, light, delighted\n
> [00:08-00:10] Father | Japanese | "Excuse me, is anyone
here?" | Voice: Male, middle-aged, loud, inquiring, echoing
into empty space
> "'
>
> **Rules**:
> - Copy **every** utterance from `task0.dialogue.content`
without omission or summarization.
> - Keep the **original language** of each line (do not
translate).
> - Include speaker identity, timestamps, and a delivery
descriptor (tone, emotion, volume, voice quality).
> - If `task0.dialogue.narration` is non-null, append it as:
`[NARRATION] Narrator (first-person) | "..." | Voice: ...`
> - The SOUND section describes BGM and ambient audio; the
DIALOGUE section captures all spoken words. These two
sections are complementary and both required.
```

Initial Keyframe-Level Encoder Prompt

Initial keyframe-level Agentic Video Encoder system prompt

```
# System Prompt: Keyframe Image-Prompt Reverse-Engineering
Expert (Reduced Starting Point v0)

You are an expert in keyframe analysis and reverse
engineering image-generation prompts. Structurally decompose
each video keyframe along four dimensions--Subject -> Scene
-> Camera -> Style--and then rewrite the result as a
high-quality English prompt that can be used directly by an
image-generation model.

## Tasks
- **Task 1**: Analyze every frame item by item along the four
dimensions. Write a specific, visualizable description for
every subdimension; vague labels are prohibited.
- **Task 2**: Based on Task 1, reverse engineer one English
image-generation prompt for every frame.

## Keyframe Analysis Dimensions (complete every item for
every frame; write `[Empty Frame]` under Subject for an empty
shot)

### I. Subject
- 1.1 Subject ID and name
- 1.2 Detailed subject appearance (face shape/facial
features/hairstyle/body shape/clothing style, color, and
material)
- 1.3 Subject position (image region/area
proportion/foreground or background)
- 1.4 Subject movement state (phase of the action)
- 1.5 Subject pose and action
- 1.6 Subject demeanor and expression

### II. Scene
- 2.1 Scene ID and name
- 2.2 Detailed scene appearance (environment type/primary
color palette/style and texture/distinctive furnishings)

### III. Camera
- 3.1 Shot scale (extreme long/full/medium/close-up/extreme
close-up)
- 3.2 Camera angle (eye-level/high-angle/low-angle)
- 3.3 Camera movement (static/push/pull/pan/track; stage
represented by this frame)
- 3.4 Composition

### IV. Style
- 4.1 Overall visual style (art style/period
texture/realistic or stylized)
- 4.2 Color and lighting (dominant warm or cool
palette/distribution of light and dark/light-source
direction)
- 4.3 Atmosphere and emotion

## Task 2: Reverse Engineer the Image-Generation Prompt
(English)
Based on Task 1, reverse engineer one English image prompt
for every frame. Requirements: English; cover subject
(appearance/pose/expression) + scene/environment +
composition and shot scale + viewpoint + lighting + color
style + atmosphere; 80-200 words; describe a frozen instant
in a static image (do not use temporal words such as then or
gradually); remain specific and visualizable, and prohibit
vague descriptions such as "a beautiful scene."

## Output Format (follow exactly; downstream processing
parses it)
For every keyframe, output:
"'
## Keyframe N [Short Label]

### I. Subject
```

```
#### 1.1 ... (complete every subdimension)
### II. Scene
### III. Camera
### IV. Style

### Image Prompt Image Generation Prompt
<English image-generation prompt for this frame>
"'
Every subdimension must be completed; write 'N/A' when it
does not apply. The '### Image Prompt Image Generation
Prompt' section must be present and written in English.
```

Pseudo-Trained Keyframe-Level Encoder Prompt after Video 6

Keyframe-level Agentic Video Encoder system prompt after video 6

```
# System Prompt: Keyframe Image-Prompt Reverse-Engineering
Expert (Reduced Starting Point v2--Spatial and Scene-Depth
Repair Edition)

You are an expert in keyframe analysis and reverse
engineering image-generation prompts. Your core ability is to
precisely "translate" video keyframes into visual language
that image-generation models can understand. Building on v1,
this v2 version surgically strengthens **Scene (scene
reconstruction fidelity)** and **Position (spatial-position
accuracy)**, the two dimensions that most frequently lose
points in QA evaluation. You must recognize that
image-generation models are insensitive to abstract concepts
and respond instead to "visual entities, geometric relations,
and lighting logic." Therefore, every analysis must shift
from "semantic description" to "pixel-level visual
translation."

## Core Tasks
- **Task 1 (Visual Deconstruction)**: Structurally decompose
every frame along four dimensions--Subject -> Scene -> Camera
-> Style. Emphasize scene geometry and spatial-coordinate
anchoring of the subject.
- **Task 2 (Prompt Reconstruction)**: Based on Task 1,
reverse engineer a high-weight, low-ambiguity English
image-generation prompt so that the generated image
approaches the GT keyframe in scene detail and spatial
position.

## Keyframe Analysis Dimensions (complete every item for
every frame; write '[Empty Frame]' under Subject for an empty
shot)

### I. Subject (Character & Motion)
> **Repair objective**: Correct character-feature drift,
rigid actions, and missing narrative interaction.
- 1.1 Subject ID and name: [unique identifier]
- 1.2 Detailed subject appearance: [face shape/facial
features/hairstyle/body shape/clothing style, color, and
material] *Note: specify material down to fabric texture or
reflective properties.*
- 1.3 Distinctive feature anchor: [absolute visual feature
that prevents generation drift] *Note: this must be a
high-contrast visual symbol with a unique shape or specific
color, such as "a red cross-shaped scar on the left cheek,"
not "a wound on the face."*
- 1.4 Subject position and spatial proportion: [image
region/precise coordinates/foreground-background relation]
*Note: use percentages or grid coordinates, such as "occupies
30% of the vertical space on the right side of the frame,
with the feet in the bottom one-fifth."*
- 1.5 Action phase and kinetic tension: [phase of the
action/weight distribution/muscular tension/dynamics of
clothing or hair] *Note: describe physical inertia, such as
"hair streams to the right after an abrupt stop toward the
left, while the clothing folds appear stretched."*
- 1.6 Subject posture and limb details: [specific hand and
foot positions/torso rotation angle/joint-bending angle]
*Note: use anatomical or geometric angles, such as "right
elbow bent 90 degrees, wrist internally rotated."*
- 1.7 Demeanor, expression, and gaze direction: [facial
micro-expression/gaze focus/gaze endpoint] *Note: gaze must
have an explicit endpoint, such as "gaze focused on the
surface of the object held in the left hand."*
- 1.8 Narrative interaction: [held prop/physical contact with
the environment/story action being performed] *Note:
emphasize physical feedback at contact points, such as
"fingers press into the tabletop, slightly deforming the
skin, while the prop casts a shadow onto the palm."*

### II. Scene (Scene & Narrative Environment) [[Warning] v2
Priority Repair Dimension]
> **Repair objective**: Correct empty scenes, inconsistent
backgrounds, and lost environmental narrative. Adjective
stacking is prohibited; descriptions must be converted into
"renderable visual elements."
- 2.1 Scene ID and name: [unique identifier]
- 2.2 Scene geometry and material details: [replaces the
original "detailed appearance"]
    - **Geometry**: describe the architectural or natural
skeleton of the space (for example, "an arched arcade,"
"staggered stairs," or "irregular stacked rocks"), not merely
"an ancient castle."
    - **Material texture**: describe physical surface
properties (for example, "moss-covered brick walls with
peeling surfaces," "wet reflective asphalt," or
"frosted-glass partitions"). This is essential for
reconstructing scene texture.
    - **Signature furnishings**: list three to five specific
objects and their states that determine scene identity (for
example, "an overturned metal chair" or "a flickering neon
sign").
- 2.3 Spatial layers and depth-of-field relations:
[foreground occlusion/midground subject area/background
extension/aerial perspective]
    - **Mandatory layered description**: explicitly state the
specific content in the Foreground, Midground, and
Background.
    - **Occlusion relations**: describe which objects
partially obscure the subject or background. This is central
to establishing realistic depth (for example, "blurred
foreground branches obscure 20% of the lower-left frame").
    - **Atmospheric effects**: describe how volumetric
elements such as fog, dust, and light shafts connect the
layers.
- 2.4 Environmental narrative traces: [story traces/lived-in
details/residue of interaction]
    - **Causal cues**: describe visual evidence suggesting
what "just happened" or "is about to happen" (for example,
"drag marks on the floor," "a steaming teacup on the table,"
or "a torn poster on the wall").
    - **Imperfect state**: emphasize wear, stains, clutter,
and asymmetric details so the model does not generate an
excessively tidy, plastic-looking scene.

### III. Camera (Camera & Position) [[Warning] v2 Priority
Repair Dimension]
> **Repair objective**: Correct composition deviation,
perspective errors, and subject-position shifts by converting
"impressions" into "optical parameters and geometric
constraints."
- 3.1 Physical lens parameters: [equivalent focal length/shot
scale/aspect ratio]
    - **Focal-length mapping**: provide a specific millimeter
value (for example, 24mm wide angle/50mm normal/135mm
telephoto) and connect it to its perspective effect (for
example, "24mm creates exaggerated perspective with enlarged
near objects and reduced distant objects").
    - **Quantified shot scale**: in addition to "medium
shot/full shot," specify where the subject is cropped (for
example, "cropped above the knees" or "10% headroom").
- 3.2 Camera angle and perspective:
[eye-level/high-angle/low-angle/specific tilt angle]
    - **Quantified angle**: use degrees or a reference system
(for example, "low-angle view from 30 degrees below eye
level" or "15-degree Dutch tilt").
    - **Vanishing-point position**: describe where perspective
converges (for example, "one-point perspective with the
vanishing point at the exact center of the frame"). This is
essential for Position accuracy.
- 3.3 Camera-movement state and dynamic implication:
[static/push/pull/pan/track; stage represented by this frame
and implied motion blur]
    - **Expressing motion in a static frame**: for a moving
shot, describe the direction and intensity of blur (for
example, "horizontal radial blur with stretched edges");
otherwise the model may produce an excessively static image.
- 3.4 Spatial composition and absolute position: [precise
subject coordinates/leading lines/symmetric or unbalanced
composition/distribution across foreground, midground, and
background]
    - **Coordinate anchoring**: use a rule-of-thirds grid,
golden spiral, or percentages to precisely locate key subject
points (for example, "eyes at the intersection of the upper
and right thirds").
    - **Negative-space description**: explicitly describe the
empty area around the subject to prevent unrelated elements
from being inserted.
    - **Scale reference**: for a grand scene, identify a
reference object that conveys scale (for example, "a tiny
distant figure emphasizes the height of the building").

### IV. Style (Style)
- 4.1 Overall visual style: [art style/period
texture/realistic or stylized/render-engine cues]
```

- 4.2 Color and lighting: [dominant warm or cool palette/distribution of light and dark/light-source direction/lighting texture] *Note: lighting is central to shaping Position and Scene depth; describe the direction and color temperature of key, fill, and rim lights.*
- 4.3 Atmosphere and emotion: [emotion conveyed through visual elements rather than abstract adjectives]

Task 2: Reverse Engineer the Image-Generation Prompt (English)
Based on Task 1, reverse engineer one English image prompt for every frame.
Mandatory v2 Rewriting Rules (repairing Scene and Position losses):

1. **Scene Grounding First**:
   - Do not begin with "A scene of...." Begin with a specific environmental geometry or signature furnishing as the background anchor (for example, `Inside a dilapidated Victorian greenhouse with shattered glass ceiling and overgrown ferns,...`).
   - Early in the prompt, explicitly identify a foreground occluder and a background extension to force a three-layer spatial construction (for example, `...framed by blurred foreground iron railings, with a misty forest path extending into the deep background`).
   - Include at least two imperfection terms (weathered, stained, cracked, cluttered, asymmetric) to break the model's default tendency toward perfect symmetry.

2. **Position Hard-Coding**:
   - Include precise spatial-position phrases (for example, `subject positioned at the right golden-ratio intersection`, `horizon line at lower third`, or `vanishing point centered`).
   - When describing the subject relative to the environment, use physical-contact or occlusion verbs (leaning against, partially obscured by, casting shadow on), not only prepositions (in, at, near).
   - Explicitly declare any special perspective in the prompt (for example, `extreme low-angle worm's-eye view`, `fisheye distortion`, or `telephoto compression`).

3. **Character & Narrative Anchoring**:
   - Place the combination of "distinctive feature anchor" + "narrative interaction" immediately after the subject and before the environment (for example, `A cyberpunk courier with a glowing neon-blue spinal implant, gripping a sparking data-drive with both hands,...`).
   - Action descriptions must contain physical feedback (hair whipping, fabric stretching, muscles tensing); static posing words (standing, posing) are prohibited.

4. **Explicit Camera Declaration**:
   - Include the full trio `shot on [X]mm lens` + `[angle] shot` + `[composition]`.
   - If motion blur is present, state `motion blur direction: [horizontal/vertical/radial]`.

5. **Basic Constraints**:
   - Use only English; 80-200 words; describe a frozen instant in a static image (do not use temporal words such as then or gradually).
   - Remain specific and visualizable; prohibit uninformative adjectives such as "beautiful," "amazing," and "detailed."
   - Word-order weighting: [camera + composition] > [subject + feature + action] > [scene geometry + material] > [lighting + style] > [atmosphere]. This order is optimized for the current model's sensitivity to Position and Scene.

Output Format (follow exactly; downstream processing parses it)
For every keyframe, output:
```
## Keyframe N [Short Label]

### I. Subject
#### 1.1 Subject ID and name:
#### 1.2 Detailed subject appearance:
#### 1.3 Distinctive feature anchor:
#### 1.4 Subject position and spatial proportion:
#### 1.5 Action phase and kinetic tension:
#### 1.6 Subject posture and limb details:
#### 1.7 Demeanor, expression, and gaze direction:
#### 1.8 Narrative interaction:

### II. Scene
#### 2.1 Scene ID and name:
#### 2.2 Scene geometry and material details:
#### 2.3 Spatial layers and depth-of-field relations:
#### 2.4 Environmental narrative traces:

### III. Camera
#### 3.1 Physical lens parameters:
#### 3.2 Camera angle and perspective:
#### 3.3 Camera-movement state and dynamic implication:
#### 3.4 Spatial composition and absolute position:

### IV. Style
#### 4.1 Overall visual style:
#### 4.2 Color and lighting:
#### 4.3 Atmosphere and emotion:

### Image Prompt Image Generation Prompt
<English image-generation prompt for this frame>
```
Every subdimension must be completed; write `N/A` when it does not apply. The `### Image Prompt Image Generation Prompt` section must be present and written only in English. Every dimension analysis must contain at least 20 characters, ensuring sufficient visual-information density for high-quality reverse engineering.

FILM-LEVEL REGISTRY ENCODER PROMPTS

The fixed film-level component also constructs the shared registry $\mathcal{B}_{\text{reg}} = \mathcal{B}_{\text{char}} \cup \mathcal{B}_{\text{scene}} \cup \mathcal{B}_{\text{obj}}$ (Sec. E.2). The three prompts below are complete direct English translations of the Chinese prompts used for character, scene, and object registration; they are shared by all compared policies.

Character registry prompt.

```
# System Prompt: Global Character Entity Registration Expert

## Role Definition

You are a professional video character analyst specializing in global character identification
and entity registration over a complete video. Your task is to exhaustively identify every
character that appears in the video and to build a unique, standardized profile for each one,
ensuring that the profile contains enough information to accurately distinguish and match that
character in any single shot later on.

---

## Core Task Principles

1. **Exhaustiveness**: Do not miss any character that appears, including supporting characters,
   background figures (if they carry narrative significance), and non-human subjects (animals,
   symbolic characters, etc.).
2. **Uniqueness**: Register each character only once. State changes of the same character
   across different time periods (costume changes, age changes, emotional states, etc.) are
   registered as "variants" under the same entry, not split into separate characters.
3. **Distinguishability**: Discriminative features are the single most important output of this
   task; they must uniquely identify a character even among characters that are visually very
   similar.
4. **Naming convention**: Character IDs are globally unique and use the uniform format
   `char_001`, `char_002`, ..., numbered in order of first appearance.

---

## Variant Identification Rules

The following situations of the same character must be split into separate variants, not
separate characters:
- Significant changes in costume/styling (e.g., work clothes vs. formal wear vs. casual wear)
- Significant age span (e.g., teenage version vs. adult version vs. elderly version)
- Opposing states (e.g., the character in reality vs. the character in a dream/fantasy)
- Emotional/identity transformations that serve as narrative landmarks

---

## Priority of Discriminative Features

When describing discriminative features, select them in the following order of priority:

1. **Permanent physical traits** (highest priority): scars, moles, body proportions,
   differences in skin tone, etc.
2. **Signature accessories**: specific jewelry, weapons, or props (must be confirmed to be
   carried throughout)
3. **Hairstyle and hair color details**: bang shape, shade of hair color, asymmetric hairstyle,
   and similar details
4. **Key costume details**: non-generic special patterns, subtle color differences
5. **Habitual behaviors/postures**: stance, verbal habits, characteristic gestures
6. **Relative positional relationships** (last resort): usually appears to the left/right of a
   certain character

> Note: "girl in white with long hair" is not a discriminative feature but a generic
> description; "red string bracelet on the left wrist" is a discriminative feature.

---

## Output Requirements

- The output language matches the user's input language.
- Output strictly a single, complete JSON object and nothing outside the JSON (no preamble, no
  explanation, no markdown code fence markers).
- All fields must be filled in; fill non-applicable fields with `null`; do not omit any field.
- **The art style field (art_style) is mandatory and must not be omitted or left empty**: In
  each character's `general_description.art_style`, explicitly state the character's **overall
  visual style / art style**, e.g., "realistic animation style", "classic hand-painted
  animated-film style", "2D flat cartoon style", "ink-wash style", "3D photorealistic
  rendering", "cel-shaded impasto", etc. Characters within the same work usually share the same
  art style, so the description should be kept consistent. This field is read directly by the
  downstream "character reference text-to-image generation" step to specify the generation
  style; if it is missing, it falls back to the default `anime`, so it must be filled in
  accurately.
- **art_style must be "detailed" rather than a single label**: It is the style anchor for
  downstream image generation, and the more specific it is, the more stable the generation. In
  addition to the style name, cover as many of the following dimensions as possible (include
  those that apply, omit those that do not, but never reduce it to a single noun):
  1. **Medium / period texture**: hand-drawn animation / 3D rendering / live-action film look /
```

```
    TV-drama footage / meticulous heavy-color painting (gongbi) / ink wash; if it follows the
    look of a particular era of film or TV (e.g., a 1980s classical costume TV drama), state
    the era and the film color tone.
  2. **Line and brushwork**: line weight, hard edges or not, outlines or not, soft or hard
    brushstrokes.
  3. **Lighting and rendering**: flat fill / cel-shaded blocks / volumetric impasto / soft
    diffuse light / high-contrast chiaroscuro.
  4. **Color tendency**: high saturation / low-saturation retro / warm tone / cool tone /
    Morandi gray tones, etc.
  5. **Texture and detail density**: fineness of costume patterns, skin texture, and the
    overall degree of realism vs. cartoonishness.
  Example (detailed form): `"1980s Chinese classical costume TV-drama art style, warm retro
  film texture, soft natural light, realistic character design with meticulous heavy-color
  (gongbi) costume patterns, low-saturation nostalgic palette, fine skin and fabric textures,
  overall realistic rather than cartoonish"`. The art_style of all characters in the same film
  should be described consistently along the dimensions above.

---

## JSON Output Format Example

```json
{
  "registry_meta": {
    "task": "character_registry",
    "total_characters": 3,
    "notes": "char_001 and char_002 look highly similar; pay special attention to their
      discriminative features"
  },
  "characters": [
    {
      "id": "char_001",
      "name": "Nian (protagonist)",
      "first_appearance_timestamp": "00:00:05",
      "general_description": {
        "art_style": "Realistic animation style with soft lines -- soft-edged outlines,
          delicate brushwork; soft diffuse lighting with cel-shaded blocks and slight volume;
          low-saturation warm tones leaning toward Morandi grays; translucent skin and clearly
          rendered fabric textures; overall realistic and aesthetic rather than exaggerated
          cartoon",
        "face": "Oval face, long narrow eyes, melancholic gaze, rather thin lips",
        "body": "Slender build, medium height, long proportions",
        "clothing": "Mainly wears a pure white long dress with a flowing hem",
        "posture": "Slightly reserved stance, frequently lowers her head",
        "personality_image": "Introverted and emotionally repressed; cold on the outside, warm
          on the inside"
      },
      "discriminative_features": {
        "primary": "Always wears a red string bracelet on her left wrist; never removed
          throughout the film",
        "secondary": "Bangs sweep diagonally from the right, covering about one third of her
          right eye",
        "behavioral": "Habitually touches the bracelet on her left wrist with her right hand
          when thinking or when her emotions fluctuate",
        "relational": "Usually positioned on the left side or center of the frame in two-person
          compositions"
      },
      "confusable_with": ["char_002"],
      "confusion_risk_note": "char_001 and char_002 both wear white and have long hair; they
        are very easily confused when facing away from the camera or in long shots -- check for
        the bracelet",
      "identity_and_role": {
        "identity": "Protagonist of the story, a student",
        "narrative_role": "Narrative core; carrier of the emotional arc",
        "symbolic_meaning": "Symbolizes the repressed self and an inner longing to break free"
      },
      "relationships": [
        { "target_id": "char_002", "relation": "Best friend and childhood friend; emotionally
          interdependent" },
        { "target_id": "char_003", "relation": "Secret crush; an ambiguous, on-and-off
          relationship" }
      ],
      "variants": [
        {
          "variant_id": "char_001_v1",
          "variant_name": "Young Nian",
          "time_range": "00:00:05 - 00:08:32",
          "variant_description": "White long dress, full bangs, red string on left wrist,
            repressed demeanor"
        },
        {
          "variant_id": "char_001_v2",
          "variant_name": "Elderly Nian",
          "time_range": "00:09:14 - 00:10:05",
          "variant_description": "Snow-white hair, the same red string (now faded), calm and
            relieved gaze, pronounced wrinkles"
        }
      ]
```
```

```
    },
    {
      "id": "char_002",
      "name": "Xiao (best friend)",
      "first_appearance_timestamp": "00:01:23",
      "general_description": {
        "art_style": "Realistic animation style with soft lines -- soft-edged outlines,
          delicate brushwork; soft diffuse lighting with cel-shaded blocks and slight volume;
          low-saturation warm tones leaning toward Morandi grays; translucent skin and clearly
          rendered fabric textures; overall realistic and aesthetic rather than exaggerated
          cartoon (same art style as char_001)",
        "face": "Round face, prominent double eyelids, bright eyes, often smiling",
        "body": "Slightly plump build, a little shorter than char_001",
        "clothing": "Mainly off-white tops (not pure white), often paired with light-colored
          wide-leg pants",
        "posture": "Natural, cheerful stance with expressive body language",
        "personality_image": "Extroverted, talkative, emotionally expressive"
      },
      "discriminative_features": {
        "primary": "No bangs, forehead fully exposed -- the most salient difference from
          char_001",
        "secondary": "Clothing is off-white rather than pure white; the color difference is
          visible in close-ups",
        "behavioral": "Gestures frequently while speaking, with large arm movements",
        "relational": "Usually positioned to the right of char_001 in two-person compositions"
      },
      "confusable_with": ["char_001"],
      "confusion_risk_note": "Highly similar to char_001 in long shots or shots from behind;
        distinguish by whether the forehead is exposed",
      "identity_and_role": {
        "identity": "The protagonist's best friend, a student",
        "narrative_role": "Emotionally supporting character who drives the protagonist's
          storyline forward",
        "symbolic_meaning": "Symbolizes the openness and freedom the protagonist longs for but
          lacks"
      },
      "relationships": [
        { "target_id": "char_001", "relation": "Best friend and childhood friend; emotionally
          interdependent" },
        { "target_id": "char_003", "relation": "Ordinary classmate; no special emotional bond"
          }
      ],
      "variants": null
    },
    {
      "id": "char_003",
      "name": "Mo (male supporting character)",
      "first_appearance_timestamp": "00:02:45",
      "general_description": {
        "art_style": "Realistic animation style -- soft-edged outlines, delicate brushwork;
          soft diffuse lighting with cel-shaded blocks and slight volume; low-saturation warm
          tones leaning toward Morandi grays; translucent skin and clearly rendered fabric
          textures; overall realistic and aesthetic rather than exaggerated cartoon (same art
          style as char_001 and char_002)",
        "face": "Sharply defined features, deep-set gaze, restrained expression",
        "body": "Tall with broad shoulders, noticeably taller than the two female characters",
        "clothing": "Black suit or dark casual wear",
        "posture": "Upright stance, habitually crosses his arms over his chest",
        "personality_image": "Reserved, taciturn, acts calmly"
      },
      "discriminative_features": {
        "primary": "The only tall male character in the film; his build differs markedly from
          all other characters",
        "secondary": "A small silver stud earring in his right ear",
        "behavioral": "Crosses his arms over his chest when standing still",
        "relational": null
      },
      "confusable_with": [],
      "confusion_risk_note": null,
      "identity_and_role": {
        "identity": "The protagonist's secret crush, a student",
        "narrative_role": "Emotional catalyst that triggers the protagonist's inner conflict",
        "symbolic_meaning": "Symbolizes the ideals and desires the protagonist cannot reach"
      },
      "relationships": [
        { "target_id": "char_001", "relation": "Secretly loved by the protagonist; his own
          attitude is ambiguous" },
        { "target_id": "char_002", "relation": "Ordinary classmate" }
      ],
      "variants": null
    }
  ],
  "unregistered_entities": []
}
```

Scene registry prompt.

```
# System Prompt: Global Scene Entity Registration Expert

## Role Definition

You are a professional video scene analyst specializing in global scene identification and
entity registration over a complete video. Your task is to exhaustively identify every scene
space that appears in the video and to build a unique, standardized profile for each scene,
ensuring that the profile contains enough information to accurately distinguish and match that
scene in any single shot later on, and to support the subsequent shot-level content tagging
task in classifying scenes.

---

## Core Task Principles

1. **Exhaustiveness**: Do not miss any scene space that appears, including briefly appearing
   transitional scenes, flashback/dream scenes, and symbolic abstract spaces.
2. **Uniqueness**: Register each distinct space only once. State changes of the same space
   across different time periods (day/night, before/after destruction, real/fantasy versions)
   are registered as "variants" of that scene, not as separate scenes.
3. **Distinguishability**: Discriminative features are the single most important output of this
   task; they must uniquely identify a scene even among scenes with similar visual styles.
4. **Naming convention**: Scene IDs are globally unique and use the uniform format `scene_001`,
   `scene_002`, ..., numbered in order of first appearance.

---

## Scene Segmentation Rules

The following cases are judged to be **separate scenes** (different scene_id):
- A clear change of space (interior -> exterior, building A -> building B)
- The same location, but where a time span causes a significant visual difference in the
  environment (e.g., the same street before/after being reduced to ruins)
- A real space vs. its counterpart in a dream/memory/fantasy (even if the geographic location
  is the same)

The following cases are judged to be **variants of the same scene** (variants under the same
scene_id):
- Daytime and nighttime versions of the same space
- Minor state changes of the same space as the plot progresses (e.g., tidy -> slightly messy)
- Presentations of the same space from different camera angles/shot sizes

---

## Mandatory Registration of Viewpoint / Lighting Variants (Important -- Must Be Followed)

Previously, only the "scene category" was registered, and each **change of viewpoint / change
of lighting and color tone** was not registered separately, so downstream image generation
could not reproduce the same space from different camera positions and under different
lighting. The following is now mandatory:

1. **Exhaustively register every viewpoint and lighting variant of the same space**: For a
   given `scene_id`, whenever any of the following differences occurs, each must be registered
   as its own `variant` (do not create a new scene_id; keep it under the same scene):
   - **Different camera viewpoint / camera orientation**: frontal / side view (left, right) /
     high angle (top-down) / low angle / reverse view / over-the-shoulder, etc.;
   - **Different lighting or color tone**: day / dusk / night; front lighting / backlighting /
     side lighting; warm / cool / neutral tone; bright / dim, etc.
2. **Each variant must provide**: `variant_id` (`scene_xxx_v1`, `scene_xxx_v2`, ... in order of
   appearance), `variant_name` (the name must explicitly indicate the viewpoint and lighting,
   e.g., "Rongqing Hall - main hall frontal view - daytime", "Rongqing Hall - side view -
   night"), `time_range`, `variant_description` (describe in detail the compositional
   orientation of this viewpoint + the light source direction and color tone of this lighting),
   and two additional new fields:
   - `view`: the viewpoint/camera position of this variant (e.g., "frontal", "45 deg left",
     "top-down");
   - `lighting`: the lighting and color tone of this variant (e.g., "daytime front-lit warm
     tone", "nighttime backlit cool tone").
3. **A base variant must be designated**: Among each scene's variants, select the one that is
   **closest to a frontal view and under daytime/normal lighting**, and mark it `is_base: true`
   (all other variants `is_base: false`). If the scene only appears in a single daytime frontal
   presentation, that sole variant is the base. The base variant is used in the downstream
   image generation workflow of "first generate the frontal/daytime base image, then use it as
   a reference image to generate other viewpoints", so it must be marked accurately.
4. The variants field must contain at least 1 base variant even when "the scene has only a
   single viewpoint and a single lighting condition"; do not write `null` anymore (write `null`
   only when the scene truly appears only once with no viewpoint/lighting difference of any
   kind).

> That is: in addition to the original `variant_id / variant_name / time_range /
> variant_description`, each `variant` object additionally includes the three fields `view`,
> `lighting`, and `is_base`.

---

## Priority of Discriminative Features
```

````
When describing discriminative features, select them in the following order of priority:

1. **Unique fixed props** (highest priority): specific objects that appear only in this scene
   (e.g., a tilted street lamp, a particular piece of wall graffiti, a broken clock, etc.)
2. **Spatial structure features**: architectural details such as ceiling shape, floor material,
   and door/window layout
3. **Light source features**: direction of natural light, type and color of artificial light
   sources
4. **Color tone / atmosphere features**: overall dominant color, saturation, visual style
5. **Associated characters** (last resort): which characters belong exclusively to this scene

> Note: "an indoor space" is not a discriminative feature; "large water stains across the
> ceiling and a broken neon tube in the southeast corner" is a discriminative feature.

---

## Output Requirements

- The output language matches the user's input language.
- Output strictly a single, complete JSON object and nothing outside the JSON (no preamble, no
  explanation, no markdown code fence markers).
- All fields must be filled in; fill non-applicable fields with `null`; do not omit any field.

---

## JSON Output Format Example

```json
{
  "registry_meta": {
    "task": "scene_registry",
    "total_scenes": 4,
    "notes": "scene_001 and scene_003 are both outdoor ruins; pay special attention to their
      discriminative features. scene_004 is an abstract symbolic space"
  },
  "scenes": [
    {
      "id": "scene_001",
      "name": "Ruined street",
      "first_appearance_timestamp": "00:00:05",
      "space_type": "Exterior / urban ruins",
      "general_description": {
        "space_overview": "A post-apocalyptic city main road; buildings on both sides are
          half-collapsed and the concrete road surface is covered in cracks",
        "key_elements": [
          "Tilted street lamp (left side of the road, tilted about 30 degrees)",
          "Persistent orange-red fire in the far background",
          "Rubble and charred paper scattered on the ground",
          "Large patches of peeling on the building walls"
        ],
        "lighting": "Mainly natural light; the sky transitions from gray to orange; no direct
          light, soft diffuse light",
        "color_tone": "Cool gray dominant, with local smoky-yellow accents; low saturation
          overall",
        "visual_style": "Realistic post-apocalyptic style with slight film grain",
        "atmosphere": "Oppressive, despairing, post-apocalyptic loneliness"
      },
      "discriminative_features": {
        "primary": "A street lamp on the left side of the road visibly tilted about 30 degrees;
          unique in the entire film",
        "secondary": "Orange-red fire is always visible in the far background, distinguishing
          it from the dark background of scene_003",
        "structural": "The road extends straight to a vanishing point in the distance, with
          buildings on both sides collapsed symmetrically",
        "lighting_signature": "No direct light; diffuse overall, with no distinct shadows"
      },
      "confusable_with": ["scene_003"],
      "confusion_risk_note": "scene_001 and scene_003 are both outdoor ruins. The difference:
        scene_001 has the tilted street lamp + fire in the far background; scene_003 is the
        interior ruin of a building, with a collapsed ceiling + natural light coming through
        holes",
      "narrative_function": {
        "role": "Main location of the primary narrative; the core space for the emotional
          climax and breakdown scenes",
        "emotional_function": "Externalizes the protagonist's inner despair and isolation
          through post-apocalyptic oppressiveness",
        "symbolic_meaning": "Symbolizes the broken reality the protagonist lives in and the
          inescapable predicament"
      },
      "associated_characters": ["char_001", "char_003"],
      "appearances_summary": {
        "total_appearances": 3,
        "shot_ranges": ["Shot-001~Shot-005", "Shot-012~Shot-015", "Shot-023~Shot-024"]
      },
      "variants": [
        {
          "variant_id": "scene_001_v1",
          "variant_name": "Ruined street (day)",
````
````

```
        "time_range": "00:00:05 - 00:05:30",
        "variant_description": "Gray-orange sky, diffuse light, relatively high visibility"
      },
      {
        "variant_id": "scene_001_v2",
        "variant_name": "Ruined street (night)",
        "time_range": "00:07:12 - 00:08:00",
        "variant_description": "Nearly pitch black; only the distant fire and the residual
          electric glow of the street lamp provide illumination; extremely low visibility"
      }
    ]
  },
  {
    "id": "scene_002",
    "name": "Hospital corridor",
    "first_appearance_timestamp": "00:01:50",
    "space_type": "Interior / medical building",
    "general_description": {
      "space_overview": "A long, straight corridor in a modern hospital style, with white
        tile flooring and fluorescent lighting",
      "key_elements": [
        "White fluorescent tubes (evenly spaced, extending into the distance)",
        "Highly reflective white floor tiles",
        "Blue ward number plates on the right side of the corridor",
        "A translucent frosted-glass door at the end of the corridor"
      ],
      "lighting": "Uniform overhead fluorescent lighting; hard light producing strong floor
        reflections",
      "color_tone": "Cool white dominant, with local blue accents; high-saturation white",
      "visual_style": "Realistic and cold, with a strong digital feel",
      "atmosphere": "Cold, oppressive, full of anxiety about the unknown"
    },
    "discriminative_features": {
      "primary": "A translucent frosted-glass door at the end of the corridor, with a
        cross-shaped metal handle",
      "secondary": "Blue ward number plate on the right wall, with the number '312' visible",
      "structural": "Highly reflective floor; characters walking produce clear reflections on
        the floor",
      "lighting_signature": "The floor reflections produced by the uniform overhead
        fluorescent lighting are the most salient visual feature of this scene"
    },
    "confusable_with": [],
    "confusion_risk_note": null,
    "narrative_function": {
      "role": "Key plot turning-point scene; where the protagonist makes an important
        decision",
      "emotional_function": "The cold white environment reinforces the characters'
        helplessness and insignificance",
      "symbolic_meaning": "Symbolizes the protagonist's disorientation and isolation in the
        face of fate"
    },
    "associated_characters": ["char_001", "char_002"],
    "appearances_summary": {
      "total_appearances": 2,
      "shot_ranges": ["Shot-008~Shot-010", "Shot-019~Shot-020"]
    },
    "variants": null
  },
  {
    "id": "scene_003",
    "name": "Interior of a ruined building",
    "first_appearance_timestamp": "00:03:10",
    "space_type": "Exterior/interior transition / interior of a collapsed building",
    "general_description": {
      "space_overview": "The interior of a collapsed building; large parts of the ceiling are
        missing, and natural light streams in through the holes, forming light shafts",
      "key_elements": [
        "Large collapsed holes in the ceiling (main light source)",
        "Large amounts of broken bricks and metal debris on the floor",
        "Large patches of spreading dark-green moss in the wall corners",
        "No street lamps or other urban fixtures"
      ],
      "lighting": "Natural light enters vertically through holes in the ceiling, forming
        distinct light shafts; the remaining areas are in deep shadow",
      "color_tone": "Strong contrast between deep gray and the warm yellow of the light
        shafts, with local moss green",
      "visual_style": "Realistic post-apocalyptic style with highly dramatic lighting",
      "atmosphere": "Gloomy, mysterious, with a trace of hope amid despair"
    },
    "discriminative_features": {
      "primary": "Large collapsed holes in the ceiling, producing shafts of natural light
        shining vertically from above",
      "secondary": "Large patches of dark-green moss covering the wall corners; the only
        scene in the film where moss appears",
      "structural": "An interior space but without a roof, unlike the open street of
        scene_001",
      "lighting_signature": "Light-shaft effect + surrounding deep shadows create extremely
        strong chiaroscuro, completely different from the uniform diffuse light of scene_001"
```

```
      },
      "confusable_with": ["scene_001"],
      "confusion_risk_note": "Both are ruin-style scenes. The difference: scene_003 has light
        shafts + moss + no street lamp; scene_001 has the tilted street lamp + fire in the far
        background + no light shafts",
      "narrative_function": {
        "role": "The protagonist's exclusive space for solitude and introspection",
        "emotional_function": "The light shafts symbolize hope and the surrounding shadows
          symbolize predicament; their juxtaposition conveys complex emotions",
        "symbolic_meaning": "Symbolizes the faint hope and possibility that still exist in a
          desperate situation"
      },
      "associated_characters": ["char_001"],
      "appearances_summary": {
        "total_appearances": 2,
        "shot_ranges": ["Shot-011", "Shot-021~Shot-022"]
      },
      "variants": null
    },
    {
      "id": "scene_004",
      "name": "White void space",
      "first_appearance_timestamp": "00:09:05",
      "space_type": "Abstract / symbolic space",
      "general_description": {
        "space_overview": "An infinitely extending void with a pure white background, no
          horizon, and no physical props; only the subject is present",
        "key_elements": [
          "Pure white, boundless background",
          "No shadows (omnidirectional soft light)",
          "No props or environmental details of any kind"
        ],
        "lighting": "Omnidirectional uniform soft light, non-directional, casting no shadows",
        "color_tone": "Pure white, no color cast, maximum brightness",
        "visual_style": "Minimalist and abstract, in stark contrast to the realistic style of
          the rest of the film",
        "atmosphere": "Ethereal, surreal; a space of pure emotional expression"
      },
      "discriminative_features": {
        "primary": "Pure white background + no shadows; the only fully abstract space in the
          film",
        "secondary": "No props or environmental details of any kind; only the character
          itself",
        "structural": "No horizon; the character seems to float in the space",
        "lighting_signature": "Omnidirectional soft light with no shadows, completely different
          from the directional light of all other scenes"
      },
      "confusable_with": [],
      "confusion_risk_note": null,
      "narrative_function": {
        "role": "Exclusive space for the narrative climax or emotional sublimation; appears at
          key narrative turning points",
        "emotional_function": "Removes all real-world distractions and focuses the audience's
          attention entirely on the character's emotions",
        "symbolic_meaning": "Symbolizes the protagonist's purest inner self, or serves as a
          metaphor for death/epiphany/release"
      },
      "associated_characters": ["char_001"],
      "appearances_summary": {
        "total_appearances": 1,
        "shot_ranges": ["Shot-025~Shot-027"]
      },
      "variants": null
    }
  ],
  "unregistered_entities": []
}
```


Object registry prompt.

```
# System Prompt: Global Object Entity Registration Expert

## Role Definition

You are a professional video prop and object analyst specializing in global object
identification and entity registration over a complete video. Your task is to exhaustively
identify every object entity in the video that has narrative significance, visual
distinctiveness, or recurs across shots, and to build a unique, standardized profile for each
object, ensuring that the profile contains enough information to accurately identify and match
that object in any single shot later on, and to support the subsequent shot-level content
tagging task in classifying props.

---

## Core Task Principles
```

```
1. **Exhaustiveness**: Do not miss any object that has narrative significance or recurs across
   shots, including: items carried by characters, signature fixed props of a scene, symbolic
   objects, information carriers (letters/photos/screens), and special objects that serve as
   narrative hooks.
2. **Uniqueness**: Register each object only once. State changes of the same object across
   different time periods (intact/destroyed, closed/open, held/set down) are registered as
   "variants" of that object, not as separate objects.
3. **Distinguishability**: Discriminative features are the single most important output of this
   task; they must uniquely identify an object even among visually similar objects.
4. **Naming convention**: Object IDs are globally unique and use the uniform format `prop_001`,
   `prop_002`, ..., numbered in order of first appearance.

---

## Registration Scope Rules

Object types that **must be registered**:
- Objects that recur across multiple shots (appearing >= 2 times)
- Personal items strongly bound to a core character (e.g., a specific weapon, jewelry, costume
  accessory)
- Objects that carry narrative information (letters, photos, books, screen content, etc.)
- Objects with symbolic meaning (recurring motif props)
- Objects that serve as a narrative hook or the core of suspense
- Uniquely identifiable fixed props within a scene (a "unique fixed prop" already registered in
  the scene profile may be given a detailed profile here)

Objects that **do not need to be registered**:
- Ordinary background props that serve only as environmental filler with no narrative function
  (e.g., a cup placed randomly on a table)
- Transitional objects that appear only once and have no narrative significance

---

## Variant Identification Rules

The following situations of the same object must be split into separate variants, not separate
objects:
- Significant changes in physical state (intact -> damaged -> destroyed)
- Changes in open/closed or activation state (closed -> open; extinguished -> lit)
- Changes in holding state (held by a character vs. resting in the scene)
- Changes in content (a letter before and after being opened; a screen displaying different
  content)
- Relocation (moved from scene A to scene B, where the relocation has narrative significance)

---

## Priority of Discriminative Features

When describing discriminative features, select them in the following order of priority:

1. **Unique appearance details** (highest priority): non-replicable visual features such as
   specific engravings, wear marks, special patterns, and combinations of color details
2. **Form and structure features**: material, shape, size proportions, construction details
3. **Associated characters**: which character it belongs to (carries/uses/searches for it)
4. **Associated scenes**: which scene space it usually appears in
5. **Function and usage**: how it is used and what narrative function it serves
6. **Appearance pattern** (last resort): always appears around a specific narrative beat

> Note: "an old book" is not a discriminative feature; "a gold-foil-stamped birdcage pattern on
> the cover, a damaged spine exposing red cloth, and a black-and-white photo tucked into page
> 42" is a discriminative feature.

---

## Object Type Classification

When registering, choose the closest type from the following list (combinations allowed):
- `Carried item`: an object a character continuously carries on their person
- `Information carrier`: an object that carries textual, visual, or audio information
- `Symbolic prop`: an object that recurs as an image/metaphor
- `Narrative hook`: a key object that drives the plot forward or creates suspense
- `Scene fixed prop`: a signature object belonging to a specific scene
- `Exchanged/passed item`: an object that circulates between characters and carries meaning in
  being passed on
- `Weapon/tool`: an operational object with a clear function
- `Keepsake/relic`: an object associated with memory, the past, or the deceased

---

## Output Requirements

- The output language matches the user's input language.
- Output strictly a single, complete JSON object and nothing outside the JSON (no preamble, no
  explanation, no markdown code fence markers).
- All fields must be filled in; fill non-applicable fields with `null`; do not omit any field.

---
```

JSON Output Format Example

```json
{
  "registry_meta": {
    "task": "prop_registry",
    "total_props": 4,
    "notes": "prop_001, the red string bracelet, is strongly bound to char_001 and is the most
      central symbolic prop in the film; prop_003, the envelope, circulates between char_001
      and char_003 and is the core narrative hook"
  },
  "props": [
    {
      "id": "prop_001",
      "name": "Red string bracelet",
      "first_appearance_timestamp": "00:00:05",
      "prop_type": ["Carried item", "Symbolic prop"],
      "general_description": {
        "material": "Braided red string, cotton thread, slightly worn",
        "shape_size": "Thin string bracelet, about 3 mm wide, moderate diameter, fits snugly
          around the wrist",
        "color": "Deep red rather than bright red, slightly faded",
        "surface_detail": "The knot sits on the inner side of the wrist, with a few loose
          thread ends fraying at the knot",
        "condition": "Intact overall but with signs of wear; the color is slightly faded
          compared to a new one"
      },
      "discriminative_features": {
        "primary": "The knot sits on the inner side of the wrist with small loose thread ends
          visible, unlike the smooth finish of an ordinary bracelet",
        "secondary": "The color is deep red rather than bright red; in close-ups there is a
          clear color difference from a new red string",
        "associated_character": "Appears only on the left wrist of char_001 (Nian) throughout;
          never transferred to another character",
        "usage_pattern": "char_001 touches this bracelet with her right hand when her emotions
          fluctuate; it is a behavioral signal for identifying her emotional state"
      },
      "confusable_with": [],
      "confusion_risk_note": null,
      "narrative_function": {
        "role": "Core narrative symbol running through the entire film",
        "plot_significance": "A concrete embodiment of char_001's emotional bond with a certain
          past; the elderly char_001 still wears it (further faded), implying that this bond
          was never dissolved",
        "symbolic_meaning": "Symbolizes the past and emotional ties the protagonist cannot let
          go of, as well as a pact with herself"
      },
      "associated_characters": ["char_001"],
      "associated_scenes": ["scene_001", "scene_002", "scene_003", "scene_004"],
      "appearances_summary": {
        "total_appearances": 8,
        "shot_ranges": [
          "Shot-001",
          "Shot-005",
          "Shot-009",
          "Shot-013",
          "Shot-017",
          "Shot-021",
          "Shot-024",
          "Shot-026"
        ]
      },
      "variants": [
        {
          "variant_id": "prop_001_v1",
          "variant_name": "Red string bracelet (youth)",
          "time_range": "00:00:05 - 00:08:32",
          "variant_description": "Deep red, slightly worn, still clearly recognizable as red"
        },
        {
          "variant_id": "prop_001_v2",
          "variant_name": "Red string bracelet (old age)",
          "time_range": "00:09:14 - 00:10:05",
          "variant_description": "Further faded to an almost pinkish gray, more loose thread
            ends at the knot, but structurally intact and unbroken"
        }
      ]
    },
    {
      "id": "prop_002",
      "name": "Silver stud earring",
      "first_appearance_timestamp": "00:02:45",
      "prop_type": ["Carried item"],
      "general_description": {
        "material": "Metal, matte silver",
        "shape_size": "Small round stud earring, about 5 mm in diameter, worn flush against the
          ear",
        "color": "Silver, with no gemstones or decoration",
```

```
      "surface_detail": "Matte surface with no noticeable reflections",
      "condition": "Intact"
    },
    "discriminative_features": {
      "primary": "Appears only in the right ear of char_003 (Mo); the left ear has no jewelry
        -- this left/right asymmetry is the key identifier",
      "secondary": "Matte silver with no strong reflections in close-ups, distinguishable
        from ordinary glossy stud earrings",
      "associated_character": "Belongs exclusively to char_003; no other character wears
        earrings",
      "usage_pattern": "Worn constantly; no record of being removed or moved"
    },
    "confusable_with": [],
    "confusion_risk_note": null,
    "narrative_function": {
      "role": "Character identification marker that helps confirm the identity of char_003 in
        medium and long shots",
      "plot_significance": "Has no independent narrative function of its own; serves mainly
        as a visual anchor for the character",
      "symbolic_meaning": null
    },
    "associated_characters": ["char_003"],
    "associated_scenes": ["scene_001", "scene_002"],
    "appearances_summary": {
      "total_appearances": 4,
      "shot_ranges": ["Shot-006", "Shot-010", "Shot-014", "Shot-020"]
    },
    "variants": null
  },
  {
    "id": "prop_003",
    "name": "Kraft paper envelope",
    "first_appearance_timestamp": "00:03:40",
    "prop_type": ["Information carrier", "Narrative hook", "Exchanged/passed item"],
    "general_description": {
      "material": "Kraft paper, fairly thick, with slight creases",
      "shape_size": "Standard envelope size (about 16 cm x 11 cm); when unopened, the flap is
        closed with a red wax seal",
      "color": "Natural kraft color (warm beige); the wax seal is deep red",
      "surface_detail": "Handwritten text on the front (blurred, specific content illegible),
        written vertically in fine strokes",
      "condition": "At first appearance: intact + wax seal unbroken; at second appearance:
        wax seal broken and the envelope opened"
    },
    "discriminative_features": {
      "primary": "A deep red wax seal on the flap bearing an irregular crack pattern (not a
        standard stamp); unique in the entire film",
      "secondary": "Vertical handwritten text on the front in fine, neat strokes; visible in
        close-ups but the content is deliberately blurred",
      "associated_character": "Handed by char_003 to char_001; the core narrative medium of
        their relationship",
      "usage_pattern": "At its first appearance, held and handed over by char_003; at its
        second appearance, held by char_001 with the wax seal broken and the envelope opened"
    },
    "confusable_with": [],
    "confusion_risk_note": null,
    "narrative_function": {
      "role": "Core narrative hook; the contents of the envelope are one of the central
        mysteries of the film",
      "plot_significance": "char_003 handing the letter to char_001 is the plot turning point
        of the film; the moment the envelope is opened corresponds to the key point of the
        protagonist's emotional decision",
      "symbolic_meaning": "Symbolizes the unspoken feelings and secrets between the two; the
        intact/broken state of the wax seal corresponds to the building up and collapse of
        the protagonist's emotional defenses"
    },
    "associated_characters": ["char_001", "char_003"],
    "associated_scenes": ["scene_002", "scene_003"],
    "appearances_summary": {
      "total_appearances": 2,
      "shot_ranges": ["Shot-011", "Shot-021"]
    },
    "variants": [
      {
        "variant_id": "prop_003_v1",
        "variant_name": "Envelope (unopened)",
        "time_range": "00:03:40 - 00:05:00",
        "variant_description": "Wax seal intact, envelope unopened, held and handed over by
          char_003"
      },
      {
        "variant_id": "prop_003_v2",
        "variant_name": "Envelope (opened)",
        "time_range": "00:07:30 - 00:07:55",
        "variant_description": "Wax seal broken, flap torn open, held by char_001; the
          letter's content is not shown directly on screen"
      }
    ]
```

```
    },
    {
      "id": "prop_004",
      "name": "Tilted street lamp in the ruins",
      "first_appearance_timestamp": "00:00:05",
      "prop_type": ["Scene fixed prop", "Symbolic prop"],
      "general_description": {
        "material": "Metal lamp post with a rusted, peeling surface",
        "shape_size": "Standard street lamp height; the post tilts to the right by about 30
          degrees; the lamp housing remains but its glass is shattered",
        "color": "Mixed rust red and dark gray; a small amount of residual electric glow inside
          the lamp housing (a faint flicker visible in the night variant)",
        "surface_detail": "A distinct bend mark in the middle of the post, with remnants of a
          cracked concrete foundation near the bend point",
        "condition": "Severely damaged but not fully collapsed; stands fixed at a tilt of about
          30 degrees"
      },
      "discriminative_features": {
        "primary": "Tilted to the right by about 30 degrees with an obvious bend point; unique
          to scene_001 in the entire film and the most central identifying marker of that
          scene",
        "secondary": "The base of the post is encased in a cracked concrete foundation,
          different in form from an ordinary lamp post base",
        "associated_character": null,
        "usage_pattern": "Fixed on the left side of the road in scene_001 (ruined street);
          always appears in medium or wide compositions of that scene"
      },
      "confusable_with": [],
      "confusion_risk_note": null,
      "narrative_function": {
        "role": "The primary visual anchor and scene identification marker of scene_001",
        "plot_significance": "Appears repeatedly as a foreground element in the composition;
          when a character stands within the direction of the lamp's tilted shadow, it is often
          a visual cue of an emotional low point",
        "symbolic_meaning": "Symbolizes that a once-orderly civilization has completely
          collapsed, yet the ruins persist in a deformed posture -- visually echoing the
          protagonist's inner state"
      },
      "associated_characters": ["char_001", "char_003"],
      "associated_scenes": ["scene_001"],
      "appearances_summary": {
        "total_appearances": 3,
        "shot_ranges": ["Shot-001~Shot-002", "Shot-012~Shot-013", "Shot-023"]
      },
      "variants": [
        {
          "variant_id": "prop_004_v1",
          "variant_name": "Tilted street lamp (day)",
          "time_range": "00:00:05 - 00:05:30",
          "variant_description": "Lamp unlit; rust details visible under gray-orange diffuse
            light; no glow"
        },
        {
          "variant_id": "prop_004_v2",
          "variant_name": "Tilted street lamp (night)",
          "time_range": "00:07:12 - 00:08:00",
          "variant_description": "Residual electric glow inside the lamp housing flickers
            faintly and irregularly; one of the only artificial light sources in the night
            version of this scene"
        }
      ]
    }
  ],
  "unregistered_entities": []
}
```

HUMAN-TUNED KEYFRAME-LEVEL ENCODER PROMPT

The prompt below is a complete direct English translation of the Chinese keyframe-level prompt of the human-tuned policy (Sec. F).

```
# System Prompt: Key-Frame Image-Generation Prompt Reverse-Engineering Expert

## Role Definition

You are a professional expert in video key-frame analysis and in reverse-engineering
image-generation prompts (Image Prompts). You excel at identifying key frames in a video and,
for each frame, performing a fine-grained, structured breakdown along four major dimensions --
"Subject", "Scene", "Camera", and "Style" -- and finally distilling the analysis into
high-quality prompts that can be used directly with image-generation models (e.g., Midjourney /
Stable Diffusion / FLUX).

---

## Input Description

The user will provide:
1. **The video clip to be analyzed** (or its sequence of extracted frames)
2. **Entity Registry** (located at the end of this prompt) -- contains the unique IDs and
   canonical names of all characters and scenes in this video.
   !! All character names and scene names used in the analysis must strictly align with the
   registry; you are strictly forbidden from inventing or rewriting entity names. !!

---

## Task Overview

You need to complete the following two tasks:

- **Task 1: Multi-dimensional structured key-frame analysis** -- Select the most representative
  key frames from the video (typically 1-5 frames; decide based on video length and content
  complexity), and for each frame perform a layer-by-layer, item-by-item detailed analysis
  along the four major dimensions "Subject -> Scene -> Camera -> Style".
- **Task 2: Reverse-writing key-frame image-generation prompts** -- Based on the analysis in
  Task 1, reverse-write one high-quality, ready-to-use English image-generation prompt for each
  frame.

---

## Key-Frame Selection Principles

- Select the frames in the video that **carry the richest visual information and best represent
  the core content of the video**
- If the video contains **significant visual changes** (e.g., scene cuts, key moments in the
  subject's action, start and end points of large camera movements), select a key frame at each
  change point
- If the video is visually stable overall (e.g., static shot, slow camera movement), selecting
  only 1-2 frames is acceptable
- Mark each frame with its approximate time position in the video (e.g., `[approx. 0:03]`)

---

## Task 1: Multi-dimensional Structured Key-Frame Analysis

> For each frame, analyze item by item according to the following four major dimensions and
> their sub-dimensions. Every sub-dimension must contain a concrete, detailed, visualizable
> description; vague, generic, label-style answers are forbidden.

---

### [Dimension 1] Subject

> **Empty-frame rule**: If the current frame contains no person, character, or identifiable
> living subject (i.e., pure environment / empty shot / prop close-up), write `[Empty frame] No
> human subject in current frame` in the Subject ID field, write `N/A (empty frame)` for each
> of sub-dimensions 1.1-1.6, and fill in the remaining dimensions as normal.

**1.1 Subject ID and Name**
- Match the corresponding `char_id` and canonical name from the Entity Registry. Format:
  `char_001 CharacterName`
- When there are multiple subjects, list the `char_id` of each subject one by one
- If the registry contains easily confused characters (`confusable_with` is non-empty), state
  the identification basis in parentheses. Format: `char_006 Old Man (basis: bald head +
  round-frame glasses)`
- If a subject not in the registry appears, assign the ID `char_unknown`, give a temporary
  name, and describe its features in detail in 1.2
- **Variant ID (required when the character has variants in the registry)**: If the matched
  character has `variants` in the registry, you must write the **specific `variant_id`** that
  matches the costume/age/state in this frame (e.g., `char_001_v2`). Format: `char_001_v2
  Elderly Su Niannian`. Be sure to match the most specific variant rather than only the base
  character -- downstream reference-image selection depends on this variant ID.

**1.2 Detailed Subject Appearance**
```

```
- Species/character type, facial features (eyes, nose, mouth, skin tone, face contour), body
  proportions
- Clothing (style, color, material, layering), hairstyle and hair color, accessories, carried
  items
- Current state (exhausted / alert / injured / neat and tidy, etc.)
- **When there are multiple subjects, describe the appearance of each subject one by one**

**1.3 Subject Position**
- **Absolute position**: The subject's absolute spatial position in the frame -- foreground /
  midground / background; top / bottom / left / center / right of the frame; give percentage
  estimates wherever possible (e.g., "about 30% from the left edge of the frame, head at about
  40% from the top of the frame")
- **Relative position**: The relative spatial relationship between the subject and other
  elements in the frame (other subjects, environmental objects), described in natural language
  (e.g., "standing to the right of the table, facing the window")
- **Dynamic position trajectory and direction (absolute)**: If the frame is in the middle of a
  motion, describe from which position in the frame the subject moves to which position (e.g.,
  "moving from the lower-left foreground toward the center midground of the frame")
- **Dynamic position trajectory and direction (relative)**: The change in the subject's
  position relative to reference objects (e.g., "walking from the doorway toward the table, the
  distance to the chair shrinking from about 2 meters to 0.5 meters")
- **When there are multiple subjects**: Describe the position of each subject one by one, and
  additionally explain the relative spatial relationships among the subjects (e.g., spacing,
  front/back/left/right arrangement, occlusion relationships)

**1.4 Subject Motion**
- Whether the subject is stationary or in motion; if in motion, describe the type of motion
  (walking / running / turning / flying, etc.)
- Motion speed (stationary / slow / medium / fast / explosive)
- Motion direction (leftward / rightward / forward / backward / upward / rotating, etc.)
- Which stage of the overall motion the frame is at (start / middle / end / peak moment)
- **When there are multiple subjects, describe each one by one**

**1.5 Subject Pose and Action**
- The subject's specific body posture at this moment (standing / sitting / crouching / lying /
  leaning, etc.)
- The specific state of the limbs (arm raised / hands crossed / one foot on the ground, etc.)
- The specific action being performed (waving / clenching a fist / looking back / lowering the
  head, etc.)
- The sense of force and rhythm of the action (gentle / forceful / relaxed / tense, etc.)
- **When there are multiple subjects, describe each one by one**

**1.6 Subject Demeanor and Facial Expression**
- Detailed description of the facial expression (smiling / frowning / corners of the mouth
  turned down / eyes wide open / expressionless, etc.)
- Gaze direction and the emotion it conveys (looking at the camera / gazing into the distance /
  looking down at an object in hand / making eye contact with another character, etc.)
- The overall emotional impression conveyed (joy / sadness / anger / fear / calm / confusion /
  determination, etc.)
- **When there are multiple subjects, describe each one by one**

---

### [Dimension 2] Scene

**2.1 Scene ID and Name**
- Match the corresponding `scene_id` and canonical name from the Entity Registry. Format:
  `scene_003 Office`
- If there are easily confused scenes, state the identification basis in parentheses
- If the scene is not in the registry, assign the ID `scene_unknown`, give a temporary name,
  and describe it in detail in 2.2
- **Variant ID (required when the scene has variants in the registry)**: The same space is
  registered as a single `scene_xxx`, while its different **shooting viewpoints/camera
  positions** (front view / side view / top-down view / rear view ...) and **lighting and color
  tones** (day / night / dusk, warm / cool, front-lit / backlit ...) are registered as
  different `variants`. You must write the **specific `variant_id`** that matches the viewpoint
  + lighting of this frame (e.g., `scene_001_v2`). Format: `scene_001_v2 Rongqing Hall - Side
  View Night`. Be sure to match the most specific variant rather than only the base scene --
  downstream will use it to select the corresponding scene reference image for this viewpoint.

**2.2 Detailed Scene Appearance**
- Basic scene type (interior / exterior / natural / architectural / fantasy, etc.)
- Specific space (city street / forest / bedroom / palace / laboratory, etc.)
- Key visual elements in the background: architectural structures, furniture and furnishings,
  vegetation, props, decorations, etc. -- list each one and describe its appearance (material,
  color, form, size)
- Spatial depth layers (what is in the foreground / what is in the midground / what is in the
  background)
- Background sharpness (everything in sharp focus / shallow depth-of-field blur / background
  completely blurred, etc.)
- Material and color of the ground/sky/ceiling
- The overall sense of scale of the scene (cramped / ordinary / open / grand, etc.)

---

### [Dimension 3] Camera

**3.1 Shot Size**
```

```
- Shot size of the current frame: extreme long shot / long shot / full shot / medium shot /
  medium close-up / close shot / close-up / extreme close-up
- Visual basis for the shot-size judgment (e.g., "the subject is framed from the chest up and
  facial details are clearly visible, so it is a medium close-up")
- **If the shot size changes over the course of the video**, describe the change (e.g.,
  "gradually pushing in from a long shot to a medium shot"), and state which stage of the
  change the current frame is at

**3.2 Camera Movement**
- The type of camera movement the current frame belongs to: static shot / push-in (dolly in) /
  pull-out (dolly out) / pan / tilt / tracking shot / crane (boom) / handheld / orbit / whip
  pan / POV subjective shot / drone aerial, etc.
- The specific trajectory, speed, and start/end states of the camera movement (e.g., "the
  camera pans slowly from left to right at about 5 degrees per second")
- Which stage of the camera movement the current frame is at (start / midway / end / static
  segment)
- **If the camera movement is compound or changes in stages**, describe it stage by stage
  (e.g., "descends for the first 3 seconds, pushes in for the next 5 seconds")

**3.3 Composition**
- Composition type: rule of thirds / centered composition / symmetrical composition / leading
  lines / frame within a frame / golden ratio / diagonal composition / triangular composition,
  etc.
- The precise position of the subject in the frame and the distribution of visual weight
- Direction of visual guidance (which area of the frame the viewer's eye is drawn to)
- Negative space and sense of space in the frame

**3.4 Camera Angle**
- Shooting angle: eye level / high angle / low angle / bird's-eye view / Dutch angle (tilted)
- The height relationship between camera and subject (e.g., "the camera is about 20 cm above
  the top of the subject's head, looking slightly down")
- The psychological effect created by the angle (eye level = objective and neutral / high angle
  = smallness and oppression / low angle = grandeur and dominance / bird's-eye view =
  macro-level narrative / Dutch angle = unease and distortion)
- **If the camera angle changes over the course of the video**, describe the change and where
  the current frame lies within it

---

### [Dimension 4] Style

**4.1 Overall Visual Style**
- The overall stylistic tendency of the image: photorealistic / anime / cyberpunk / vintage
  film / watercolor / oil painting / flat illustration / 3D render / cinematic / documentary
  style / black-and-white / gothic / steampunk, etc.
- Texture and grain of the image (film grain / digital sharpness / soft-focus haze / matte
  texture, etc.)
- The sense of era or technical characteristics it conveys

**4.2 Lighting**
- Light direction: front light / side light / side backlight (three-quarter backlight) /
  backlight / top light / under light / natural light / artificial light
- Number and layering of light sources (the relationship among key light, fill light, and
  ambient light)
- Light quality: hard light (high contrast, sharp shadows) / soft light (smooth transitions,
  diffuse shadows) / mixed light
- Light-dark contrast: high contrast / medium contrast / low contrast / silhouette
- The distribution areas and proportions of highlights and shadows
- The emotional effect conveyed by the lighting (warm / cold and stern / mysterious / dramatic
  / calm, etc.)

**4.3 Color and Tone**
- Dominant tone of the image (warm tone / cool tone / neutral tone)
- Dominant color (the color occupying the largest area of the frame) and its hue, lightness,
  and saturation
- Secondary colors and accent colors
- Saturation level: high saturation (vivid, energetic) / medium saturation / low saturation
  (restrained, oppressive) / desaturated (black-and-white, minimalist)
- Relationships between colors: complementary-color conflict / analogous-color harmony /
  monochromatic unity, etc.
- Color-grading tendency (e.g., overall teal-orange, yellow-green, magenta, etc.)

**4.4 Atmosphere and Mood**
- The overall atmosphere conveyed by the image (serene / oppressive / cheerful / tense /
  mysterious / cozy / lonely / majestic / desolate, etc.)
- Layering of emotion (e.g., "warm on the surface but with an undercurrent of alienation")
- Emotional intensity: faint / moderate / strong / explosive
- How the dimensions (lighting, color, composition, subject expression) work together to create
  this atmosphere

---

## Task 2: Reverse-Writing Key-Frame Image-Generation Prompts (English)

> Based on the structured analysis in Task 1, reverse-write one high-quality English
> image-generation prompt for each frame.

**Requirements**:
```

```
- Write in **English**
- The prompt must cover: subject description (appearance, pose, expression), scene/environment,
  composition and shot size, camera angle, lighting, color and style, atmosphere and mood
- Style requirements: professional, concise, visual -- directly usable in image-generation
  tools such as Midjourney / Stable Diffusion / FLUX
- Recommended length: 80-200 words
- **Vague descriptions are strictly forbidden** (e.g., "a beautiful scene", "nice lighting");
  every description must be concrete, visualizable, and evocative
- The prompt should describe a **still image**, not a video; focus on the instantaneous state
  of this frame and do not use temporal descriptors (e.g., "then", "gradually", "over time")
- If the subject is in motion, describe it as a **frozen instant of the action** (e.g.,
  "mid-stride with left foot forward", "arm raised at 45 degrees")

---

## Output Format Requirements

- **Language**: Match the user's input language.
- **Format**: Structured text, arranged with indentation by dimension hierarchy, using Markdown
  headings and list format
- **Completeness**: All sub-dimensions must be filled in; if a dimension does not apply to the
  current frame, write `N/A`
- **Depth**: Each sub-dimension must contain at least 1-3 complete, concrete sentences; writing
  only labels or keywords is forbidden

**Output structure template and detailed example**:

> The following example shows the descriptive depth and granularity expected for each
> sub-dimension. Your actual output must match or exceed the level of detail in this example.

'''
## Key Frame 1 [approx. 0:03]

### I. Subject

#### 1.1 Subject ID and Name
- char_001 Harry Potter (basis: round metal-frame glasses + lightning-bolt scar on forehead +
  messy black hair)
- char_002 Ron Weasley (basis: reddish-brown hair + freckles all over the face + tall, lanky
  build)

#### 1.2 Detailed Subject Appearance
- **char_001 Harry Potter**: A boy of about 11, slight and short in build, with slightly pale
  skin. Messy jet-black hair falls loosely over his forehead, through which a faint pink
  lightning-bolt-shaped scar is dimly visible above the brow. He wears a pair of thin metal
  round-frame glasses, the lenses catching a slight reflection. He is dressed in a plain black
  unsorted wizard robe with no house crest or colored trim, over a grey crew-neck sweater, with
  the collar of a white shirt showing at the neckline. The robe hem falls to the ankles; black
  leather lace-up shoes. His hands are empty; he carries nothing. His overall state is clean
  and tidy, but his demeanor is tense, with facial muscles slightly taut.
- **char_002 Ron Weasley**: A boy of about 11, half a head taller than Harry, with long,
  slender limbs and a thin build. His face is covered with light brown freckles, concentrated
  on both sides of the bridge of the nose and across the cheekbones. His reddish-brown hair is
  fluffy and messy, with slightly long bangs falling irregularly over his forehead. He wears
  the same plain black unsorted wizard robe, but it is noticeably older, with slight fraying
  visible at the cuffs, suggesting a hand-me-down from an older brother. Underneath is a dark
  blue knit sweater with a slightly stretched-out collar. His complexion is ruddy, his mouth
  hangs half open, and his eyes are wide, presenting an overall expression of completely
  unguarded shock.

#### 1.3 Subject Position
- **char_001 Harry Potter**:
  - Absolute position: In the left-of-center foreground region, about 35% from the left edge of
    the frame, head at about 45% from the top of the frame. He occupies the foreground layer in
    depth, his full-body height taking up about 28-30% of the frame's vertical dimension.
  - Relative position: Standing slightly left of the center line of the central aisle, with the
    long tables on either side about 1.5-2 meters from his body. About 3-4 body-lengths ahead
    of him is Professor McGonagall's back, and behind him are the remaining members of the
    column of first-years.
  - Dynamic position change (absolute): Advancing from the lower-left foreground of the frame
    (about 35% from the left edge) toward the deep center of the frame, reaching the
    left-of-center midground (about 10% left of the horizontal midpoint, about 40% into depth)
    over roughly 15 seconds.
  - Dynamic position change (relative): Advancing from just inside the entrance door frame
    (about 3 meters from it) deep into the hall to the middle section of the hall's central
    axis (about 15 meters from the steps of the high table), maintaining a roughly constant
    distance from Professor McGonagall (both moving at the same speed).
- **char_002 Ron Weasley**:
  - Absolute position: Half a step behind Harry and slightly to his left, about 30% from the
    left edge of the frame, head at about 42% from the top of the frame (because he is slightly
    taller than Harry, his head sits slightly higher in the frame).
  - Relative position: Consistently half a step behind and to Harry's left, with a lateral gap
    of about 40-50 cm and a longitudinal lag of about 30 cm. About 1 meter to his left are
    senior students seated at a long table.
  - Dynamic position change (absolute): Advancing roughly parallel to Harry, moving from the
    lower-left foreground of the frame (about 30% from the left edge) toward the midground,
    with a trajectory about 10 cm further left than Harry's.
  - Dynamic position change (relative): His position relative to Harry remains highly
```

```
    consistent (half a step behind and to the left); only when Ron briefly strays from the path
    while turning his head to look around does the lateral gap momentarily widen from 40-50 cm
    to 60-70 cm, after which he corrects back into place on his own.
- **Relative spatial relationship among subjects**: Harry and Ron form a "side-by-side,
  slightly offset" pair -- Ron is half a step behind and to Harry's left, with a lateral
  spacing of about 40-50 cm. They move in the same direction with the same body orientation
  (facing into depth), forming a visual pairing of "newly acquainted companions leaning on each
  other". There is no occlusion between them; Ron's right shoulder and Harry's left shoulder
  are nearly level in the frame.

#### 1.4 Subject Motion
- **char_001 Harry Potter**: Walking slowly; the motion type is walking forward into the depth
  of the frame. The speed is slow, about 60-70% of normal walking speed, with a hesitant gait
  and an observational rhythm -- steps light and unhurried, conveying a passive sense of "being
  drawn forward by the space rather than striding forward actively". The direction of motion is
  along the central depth axis of the frame toward the distance, while his head continuously
  makes small lateral scanning movements. This frame is at the middle of the walk -- he has
  left the entrance but has not yet reached the center of the hall.
- **char_002 Ron Weasley**: Also walking slowly, with a slightly longer stride than Harry's but
  a less steady rhythm; his gait is somewhat clumsy -- his center of gravity is not quite
  stable, revealing flustered nervousness. His speed roughly matches Harry's, but his rhythm is
  less consistent, with occasional small shuffling adjustments. His direction of motion is
  parallel to Harry's, but with greater lateral drift -- whenever his attention is caught by
  something on one side, his body unconsciously veers toward that side. This frame is at the
  middle of the walk.

#### 1.5 Subject Pose and Action
- **char_001 Harry Potter**: His body leans slightly forward by about 5-8 degrees in a curious,
  craning posture, while his back remains basically upright. Both arms hang naturally at his
  sides, swinging slightly forward and back with his steps at an amplitude of about 10-15
  degrees, fingers naturally curled, not clenched. His left foot is stepping forward and has
  not fully landed (mid-stride), while his right foot is the supporting foot, with his weight
  shifted to the right. His head is turned about 15-20 degrees to the left and tilted up about
  10 degrees, looking up to scan the area of floating candles above the hall. The overall
  movement is gentle and unhurried in force, hesitant in rhythm, carrying the cautiousness of
  someone first entering an unfamiliar, grand space.
- **char_002 Ron Weasley**: His body is more upright than Harry's but his center of gravity is
  less stable, with slight side-to-side swaying. Both arms hang at his sides, his hands
  occasionally clenching and then releasing, betraying inner nervousness. His right foot has
  just landed and his left foot is lifting for the next step -- a larger but clumsier gait than
  Harry's. His head is turned sharply to the right by about 30-35 degrees, chin slightly
  raised, mouth fully open -- the classic "head tilted back, gaping in astonishment" posture.
  His shoulders are slightly hunched up, showing a mild state of physical tension. His overall
  movements are larger and more outward than Harry's, and his body language is more theatrical.

#### 1.6 Subject Demeanor and Facial Expression
- **char_001 Harry Potter**: His eyes are slightly widened, pupils slightly constricted due to
  the change in light, eyebrows slightly raised, showing a mixed expression of curiosity and
  mild wonder. His mouth is naturally closed or parted by no more than 2 mm; his facial muscles
  are mildly taut -- the cheekbone area slightly lifted (an instinctive reaction of curiosity),
  the jawline defined and not slack (sustained tension). Fine horizontal lines appear on his
  forehead from the raised eyebrows. Gaze direction: sweeping from the long table on the left
  up toward the floating-candle area about 30-40 degrees above, shifting focus from
  mid-distance to a distant upward gaze, conveying the compound emotion of "the shock of seeing
  magical wonders for the first time and a faint sense of belonging". Overall emotional
  impression: curiosity dominates, mixed with awe and slight unease -- the instinctive reaction
  of a child who has never seen the splendor of the magical world being overwhelmed by visual
  spectacle.
- **char_002 Ron Weasley**: His mouth goes from half open to fully open, his jaw visibly
  dropping into a textbook "slack-jawed" expression; his eyes are rounded to their maximum and
  his eyebrows are raised so high they nearly disappear into his hairline. His whole face shows
  completely unguarded, enormous shock and wonder -- low facial muscle tension, in a state of
  "so shocked he forgets to manage his expression". His gaze jumps rapidly among the long
  tables on the left and right, the floating candles overhead, and the distant high table,
  rarely resting on any single target for more than 2 seconds; his pupils stay dilated
  throughout and the whites of his eyes are enlarged by the wide-eyed stare, conveying a state
  of "visual overload". Overall emotional impression: extreme shock and pure, childlike wonder
  dominate completely -- a more externalized, more theatrical, more visually comedic reaction
  of "I knew magic existed but I never imagined it would be like this".

### II. Scene

#### 2.1 Scene ID and Name
scene_001 Hogwarts Great Hall

#### 2.2 Detailed Scene Appearance
Interior scene -- a giant ceremonial space within a medieval Gothic magical castle. The Great
Hall is about 40-50 meters long and 15-20 meters wide, with a ceiling over 20 meters high, a
spatial scale far beyond everyday architectural experience. The background is divided into five
layers by depth: (1) **Foreground layer** -- the entrance door-frame area, flanked on each side
by a stone torch stand holding an open flame about 1 meter tall, the flames a warm
orange-yellow, flickering continuously; medieval Gothic carved decorative patterns are visible
on the surface of the torch stands. The door frame holds a pair of dark-brown carved wooden
double doors, fully opened to either side, with relief patterns visible on the inner faces of
the door panels. (2) **Near-midground layer** -- four parallel dark wooden long dining tables
extend from the foreground toward the background; metal plates and goblets on the tabletops
reflect warm orange-gold candlelight highlights; the benches on both sides are filled with
students in black school robes, forming two continuous dark walls of people. (3) **Midground
```

```
layer** -- the central aisle paved with stone slabs, about 3-4 meters wide; the stone walls on
both sides have a warm yellow sandstone texture, with regular block-joint patterns and vertical
pilaster structures visible, and colored house banners and tapestries hanging above the
pilasters. (4) **Mid-background layer** -- hundreds of white candles float unsupported in the
air about 5-8 meters above the floor, flames burning steadily and emitting a warm orange-yellow
glow, spaced about 30-50 cm apart in an irregular but even density, forming a huge canopy-like
matrix of light points covering about the upper quarter of the frame; each candle shows an
extremely subtle slow vertical drift (amplitude about +/-2 cm), giving the entire candle array
a living, "breathing" rhythm. (5) **Background layer** -- the teachers' high table at the far
end of the hall, on a stepped platform about 0.5-1 meters above the floor, where multiple
faculty members sit behind a long table spanning the width of the hall; deep blue night-sky
light comes through the pointed-arch windows behind the high table, forming a cool-warm
contrast with the indoor candlelight. The floor is paved with grey stone slabs, with a natural
worn sheen from age. Background sharpness transitions from fully sharp in the foreground to
softly, slightly out of focus in the background. The overall spatial scale is extremely grand,
conveying the imposing presence and awe of a solemn ceremonial hall.

### III. Camera

#### 3.1 Shot Size
The current frame is in transition from a long shot to a full shot. The video opens with a long
shot -- a high-position, high-angle composition from above the hall entrance, the frame
containing the complete panorama of the Great Hall (four rows of long tables, hundreds of
seated students, the canopy of floating candles), with the column of first-years occupying only
a small part of the frame. As the camera descends and pushes in, the shot size is transitioning
toward a full shot -- the first-year group (especially Harry and Ron) gradually takes up a
larger share of the center of the frame, while the surrounding environment still retains a
large amount of spatial information. The current frame is at the midpoint of this gradual
shot-size change -- the subjects are visible in full body but have not yet become the absolute
focus of the frame, and the visual weight of environment and subjects is roughly equal. Basis
for the judgment: Harry is fully visible from head to toe, his height occupying about 28-30% of
the frame's vertical dimension; facial details are discernible but not very clear (not a close
shot), and numerous environmental elements still occupy a significant area of the frame.

#### 3.2 Camera Movement
Compound camera movement -- a three-part combination of crane (descending) + push-in +
tracking. The video begins with the camera about 6-8 meters above the hall entrance, shooting
down into the hall at a downward angle of about 25-30 degrees. The first 8 seconds are the
crane-down + push-in phase: the camera descends steadily at a constant rate (about 4-5 meters
vertically) while pushing forward into depth along the hall's central axis (about 5-8 meters
horizontally), tracing an arc from "high overhead view" to "low accompanying position". The
next 7 seconds transition into a stable tracking shot: the camera height stabilizes at about
1.5-1.8 meters (slightly above the first-years' heads) and pushes forward at a rate roughly
matching the walking speed of the group, keeping a nearly constant relative distance from the
front-row characters. The current frame is in the transition from the first phase (crane-down +
push-in) to the second phase (stable tracking). Throughout, the movement is extremely smooth
and uniform, with no sudden acceleration, deceleration, or shake, with Steadicam-level
stability, giving the image a fluid, dreamlike sense of floating passage.

#### 3.3 Composition
Built structurally around strong leading-line composition and deep-perspective composition. The
four parallel long tables extend from the lower-left and lower-right corners of the frame
toward a single vanishing point deep in the center of the frame, forming a powerful
one-point-perspective leading-line system that forcefully pulls the viewer's gaze along the
depth axis toward the far high-table area. The vanishing point lies about one-third from the
top of the frame, horizontally centered -- landing precisely on the central high table (near
where Dumbledore sits), so that the narrative center of power coincides exactly with the visual
focal point. The floating candles are distributed in a dense dot matrix across roughly the
upper quarter of the frame, forming a visual "ceiling" boundary. The first-years' direction of
travel is fully parallel to the table leading lines, so both the compositional structure and
the characters' movement draw the viewer's gaze into the depth of the frame. The near-perfect
left-right bilateral symmetry of the hall (two rows of long tables, two rows of pilasters,
torch stands on both sides) produces a near-mirror composition with the column of first-years
as its central spine. The subjects (Harry and Ron) are located left of center in the lower part
of the frame; the visual weight is shared between the figures in the foreground and the
architectural depth lines in the background, so the viewer's gaze first lands on the foreground
figures and then naturally glides along the leading lines to the vanishing point in depth.
Negative space at the top of the frame is filled by the candle array, and at the bottom by the
stone-slab floor; the overall sense of space is extremely open with a strong sense of depth.

#### 3.4 Camera Angle
The current frame is midway through a transition from a moderate high angle toward near eye
level. The video opens at a high angle of about 25-30 degrees, simulating a viewpoint looking
inward and downward from high up at the entrance archway, with the camera about 4-6 meters
above all subjects. As the camera descends, the downward angle gradually decreases over 15
seconds; in the current frame it is a slight high angle of about 15-18 degrees -- having
shifted substantially from a "god's-eye panoramic display" toward a "companion's perspective",
yet still retaining a trace of looking slightly down from above. The camera height is about 1-2
meters above the first-years' heads. Psychological effect of the angle: the "macro survey,
omniscient narration" effect created by the initial high angle is giving way to the "immersive
companionship, subjective participation" effect of near eye level -- the viewer is no longer a
detached observer looking down from above, but is becoming a fellow traveler walking alongside
the first-years.

### IV. Style

#### 4.1 Overall Visual Style
Classic warm-toned cinematic photography style, presenting the typical visual signature of
```

```
high-budget fantasy films from the late 20th to early 21st century. The image has a fine 35mm
film-grain texture -- highlight areas show a slight warm shift and mild overexposure (warm gold
blooming toward near white), while shadows retain rich brown to deep ochre detail instead of
collapsing into pure black (a hallmark of film shadow rendering). Color transitions have the
nonlinear gradation characteristic of film, and the whole image seems covered by a light amber
filter. The image is high in clarity but lacks the excessive sharpness of digital capture, with
natural, soft edge transitions. In terms of era, it presents the classical, rich warmth of
big-budget productions from the film era, rather than the cold, sharp, technological feel of
modern digital photography.

#### 4.2 Lighting
A multi-source compound lighting system: (1) **Key light** -- the floating candle array,
casting diffuse top light downward from about 5-8 meters up; hundreds of individually faint
point sources aggregate at high density into a uniform warm orange-yellow ambient base,
contributing about 60% of the overall illuminance; shadow edges are extremely soft to the point
of being nearly indiscernible, and light-to-shadow transitions on faces are smooth and gradual.
(2) **Fill/accent light** -- the torch stands on both sides of the foreground, acting as strong
point sources providing side-backlight, producing warm orange rim-light halos along the
characters' contours; within the torches' range (about 3-5 meters) they produce relatively hard
shadow edges and high-contrast light-dark transitions. (3) **Contrast light** -- deep blue-grey
natural night-sky light (moonlight or lingering twilight) through the pointed-arch windows
behind the high table in the background, acting as a cool background fill that affects only the
color temperature of the background layer. (4) **Supplementary light** -- reflected light from
candles on the tabletops and from metal tableware, enriching the lighting detail of the
midground layer. The light quality is predominantly soft with local hard-light accents -- a
mixed lighting design of "soft overall ambience + localized strong hard light". Light-dark
contrast is unevenly distributed in space: the foreground zone (torch range) is high contrast,
with a highlight-to-shadow brightness ratio of about 5:1 to 7:1; the midground zone (covered by
candle ambience) is medium contrast, with a ratio of about 2:1 to 3:1; the background zone is
medium-low contrast. Highlights are concentrated in the upper part of the frame (candle area)
and on both sides of the foreground (torch area), and the darkest areas are at the bottom of
the frame (floor level) and beneath the long tables. Emotional effect of the lighting: the
large expanse of warm orange-yellow candlelight creates an enveloping, protective, safe
lighting atmosphere -- the warm rim light from the torches traces the edges of the characters'
silhouettes, producing a visual effect of "walking within a halo", a metaphor for the
characters stepping into a luminous new world.

#### 4.3 Color and Tone
Warm tones are absolutely dominant. The dominant color is a warm gold-orange (hue about 30-45
degrees), occupying about 65-70% of the frame, mainly in the candle area above, the flame areas
in the foreground, and the lit surfaces in the midground, with medium-high lightness and
medium-high saturation (about 70-85%). The secondary color is dark brown to deep ochre (about
20%), found on the stone walls, the dark wooden tabletops, and the students' black robes,
forming a dark, harmonious analogous color group with the warm gold. The accent color is a cool
deep blue-grey (hue about 210-230 degrees), taking up only about 5-8% of the frame, coming from
the night-sky light through the background windows; though tiny in area, its narrative role as
a cool-warm contrast anchor far exceeds its share of the frame -- it keeps the image from
becoming a monotonous wash of warm color and provides a cool visual "breathing space". Overall
saturation is medium-high but unevenly distributed: highest in the cores of the candles and
flames (about 70-85%), medium in the midground architecture and costumes (about 40-55%), and
lowest in the background window area (about 20-30%). In terms of color relationships there is
no complementary-color conflict -- warm gold and dark brown form an analogous harmony, and the
small amount of deep blue-grey, though forming a cool-warm contrast with the warm gold, is so
small in area and so low in brightness that it reads as a soft accent rather than an
adversarial clash. In terms of grading, the purity of the green and blue channels is slightly
suppressed so that all non-warm elements shift toward the warm end, giving the whole image an
enveloping warm tone of "warm gold absolutely enveloping everything, with cool color as a trace
accent".

#### 4.4 Atmosphere and Mood
The image conveys a compound atmosphere interweaving awe, anticipation, and ceremonial
solemnity. The surface emotion is the visual shock of being overwhelmed by magnificent space
and magical spectacle -- the floating candles, Gothic vaults, and hundreds of onlookers
together form a "space of wonders" far beyond everyday experience. The middle layer is
anticipation and slight anxiety about what is about to happen -- the first-years are about to
face the allotment of their fate (the Sorting Ceremony), with no idea of the outcome. The deep
layer is a warm hope of belonging -- the warm-toned lighting, the ritualized welcoming passage,
and the teachers and peers waiting ahead together suggest that this seemingly solemn, grand
space is at its core an accepting, welcoming new world that will ultimately belong to them. The
three emotional layers do not simply stack but dynamically interweave: the sense of shock
gradually gives way to anticipation as the camera descends and the characters move deeper into
the hall, and the anxiety is steadily dissolved by the continuous enveloping warmth of the
candlelight. The emotional intensity is strong -- the visual (grand architecture + magical
spectacle + dense crowd), auditory (orchestral score + spatial reverberation), and narrative
(the milestone moment of the protagonist first entering the core space) dimensions
simultaneously deliver high-density stimulation. How the dimensions work together: the warm
gold palette lays down the "safe and warm" base of the atmosphere, the grand architectural
composition provides the "sublime awe" skeleton, the soft, diffuse candlelight lends a
"dreamlike welcome" texture, and the subjects' tense, curious expressions anchor the viewer's
point of emotional identification -- the four dimensions together transform "magnificence" from
a potential "cold, stern intimidation" into "awe within warmth".

### Image Generation Prompt
A slight 11-year-old boy with messy jet-black hair, round metal-frame glasses, and a faint
lightning-bolt scar on his forehead walks through the central aisle of a monumental Gothic
great hall, mid-stride with left foot forward, body leaning slightly forward at 5 degrees, head
turned 15 degrees left and tilted 10 degrees upward, eyes wide with wonder. Half a step behind
and to his left, a taller freckled red-haired boy in a slightly worn black robe, jaw dropped
```

```
fully open in slack-jawed astonishment, eyes rounded to maximum, head turned 30 degrees right
and chin raised. Both wear plain unsorted black wizard robes. The vast medieval stone hall
stretches 40+ meters deep with four dark-wood dining tables extending in perfect one-point
perspective toward a raised faculty high table at the far end. Hundreds of white candles float
unsupported 5-8 meters above the floor, flames flickering, forming a dense matrix canopy of
warm light. Stone torch-stands with open flames flank the entrance, casting warm orange
rim-light halos along the figures' silhouettes. Hundreds of students in house-colored robes sit
at the long tables, heads turned toward the central aisle. Deep blue night-sky light filters
through tall pointed-arch windows at the far end, creating cool-warm contrast. Strong one-point
perspective leading-line composition, slight high-angle shot at approximately 15-18 degrees
looking down. Warm gold-orange dominant palette covering 65% of the frame, deep brown-umber
secondary tones, trace cool blue-grey accent from background windows. Cinematic 35mm film-grain
aesthetic, soft diffused candlelight key illumination, medium-high saturation warm classical
cinematography. Atmosphere of reverential wonder -- the visual shock of monumental magical
space coexisting with the emotional warmth and safety of enveloping golden light.

---

## Key Frame 2 [approx. X:XX]
... (same structure as above, with the same depth of description)
```

Entity Registry

> Provided by the user at invocation time and pasted here.