PIXELLE-VIDEO: A CROSS-SHOT AND CROSS-MODAL COORDINATION FRAMEWORK FOR END-TO-END VIDEO PRODUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent video generation models can synthesize visually compelling short clips, yet producing a complete video requires coordinating persistent information across shots, modalities, and heterogeneous production stages. We introduce **Pixelle-Video**, a fully automated engine for end-to-end video production. A Stateful Storyboard persists throughout production, accumulating multimodal artifacts and execution metadata as a shared evolving state across stages. Building on this, Pixelle-Video introduces two coordination mechanisms: (1) Cross-Modal Temporal Coordination, which uses the duration of synthesized narration as a shared temporal clock for visual processing and clip composition; and (2) Persistent Visual Memory, which carries character, scene, and visual references across shots and combines multimodal inspection with targeted regeneration to maintain visual consistency. Evaluations on VBench, ViStoryBench and DirectorBench show that Pixelle-Video achieves the best performance across all nine dimensions among representative automated video-production systems, spanning semantic consistency, cross-modal synchronization, and overall video quality.



Figure 1: **Overview of Pixelle-Video.** Given a high-level user prompt, Pixelle-Video produces a coherent multi-shot video while maintaining visual consistency of the orange cat wearing a red collar with a gold bell, the young woman, and the living-room environment across shots, as well as cross-modal temporal synchronization among text, video, and audio modalities.

1 INTRODUCTION

Recent advances in video generation have substantially improved the quality of synthesized videos. Models such as Sora (Brooks et al., 2024), Seedance (Bai et al., 2026), and Veo (Google DeepMind, 2025) have demonstrated strong capabilities in visual fidelity, motion realism, and prompt following, making it increasingly feasible to generate visually impressive video clips. However, complete video production requires coordinating multiple shots and heterogeneous modalities into a coherent whole, rather than generating each clip in isolation (Xu et al., 2024; Shi et al., 2025; Meng et al., 2026). A practical end-to-end video production system should be able to turn a brief idea into a complete

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.

video while maintaining coherent narrative progression across shots, synchronizing heterogeneous modalities over time, and preserving reliable production context throughout execution (Shi et al., 2026a; Luo et al., 2023).

Existing video generation pipelines, such as Anim-Director (Li et al., 2024), MM-StoryAgent (Xu et al., 2025a) and MoneyPrinterTurbo (Harry, 2024), commonly organize creation as a sequence of stages, including script planning, visual generation, audio synthesis, and final composition. Nevertheless, these systems still rely on loosely coupled stage-wise handoffs, where individual shots or modalities are produced separately and later assembled into a complete video. Existing methods still face the following major limitations. First, cross-modal temporal coordination remains challenging. Cross-modal assets are often produced under independent timing assumptions and aligned during composition, making synchronization difficult across segments (Liu et al., 2026; Song et al., 2026; Weng et al., 2026). Second, they struggle to maintain cross-shot visual consistency. Since clips are usually generated under loosely connected prompts, characters may drift in appearance, and scene layouts may change unexpectedly (Zhu et al., 2025; Yuan et al., 2024). Besides, production-level quality control often relies on repeated manual prompt engineering and visual inspection, resulting in high iteration cost and limited scalability (Long et al., 2026; Meng et al., 2026).

To address these limitations, we introduce **Pixelle-Video**, a fully automated engine for end-to-end video production. Rather than treating planning, generation, and post-production as loosely connected stages, Pixelle-Video organizes the entire workflow around a *Stateful Storyboard* that persists throughout execution. Each storyboard segment is progressively populated with narration, visual assets, timing information, intermediate results, and final clips, providing a shared representation through which heterogeneous production stages can coordinate. This design allows the storyboard to serve not only as a planning structure, but also as the evolving production context that connects upstream generation with downstream rendering and composition.

Building on this representation, Pixelle-Video introduces two key coordination mechanisms. First, *Cross-Modal Temporal Coordination* uses the duration of synthesized narration as a shared temporal reference for subsequent visual processing and clip composition. Instead of independently generating multimodal assets and reconciling their durations only at the end, Pixelle-Video propagates narration timing through the segment execution process, thereby coupling visual content and composition to the same temporal structure. Second, *Persistent Visual Memory* maintains reusable character, scene, and visual references across shots, providing subsequent generations with persistent cross-shot context. Multimodal inspection further evaluates intermediate visual results and triggers targeted regeneration when inconsistencies are detected, reducing the need for repeated manual correction while improving long-range visual consistency.

We evaluate Pixelle-Video against five representative automated video-production systems using metrics from VBench, ViStoryBench, and DirectorBench. Pixelle-Video achieves the best performance across all nine dimensions, with particularly clear gains in cross-shot coherence and cross-modal alignment. Human evaluation further confirms that these improvements are reflected in perceptually better final videos, with Pixelle-Video preferred over each competing system in 70.3%–81.2% of pairwise comparisons. Ablation studies verify the complementary contributions of Persistent Visual Memory and Cross-Modal Temporal Coordination. Together, these results suggest that persistent production state and explicit cross-shot and cross-modal coordination provide an effective basis for coherent end-to-end video production.

Our contributions are threefold:

- We present **Pixelle-Video**, a fully automated engine for end-to-end video production, built around a persistent Stateful Storyboard that connects planning, generation, rendering, and composition within a unified evolving production process.
- We introduce two coordination mechanisms: *Cross-Modal Temporal Coordination* for synchronizing heterogeneous modalities over time, and *Persistent Visual Memory* for maintaining cross-shot visual consistency with automated inspection and targeted regeneration.
- Comprehensive evaluations show that Pixelle-Video achieves the best results across all nine automatic metrics and consistent human preference over all competing systems.

2 Related Work

Coherent Multi-shot Video Generation. Recent advances in video generation have substantially improved visual quality, motion realism, and text alignment, from early diffusion-based models such as VDM (Ho et al., 2022), Stable Video Diffusion (Blattmann et al., 2023), ModelScopeT2V(Wang et al., 2023) and VideoCrafter (Chen et al., 2023; 2024), to recent video generation models such as Sora (Brooks et al., 2024), Wan (Wan et al., 2025) and Seedance (Bai et al., 2026). Beyond isolated clips, increasing attention has been devoted to temporally extended and multi-shot generation, where maintaining coherent subjects, scenes, and narrative context becomes critical (Xu et al., 2025b; Shi et al., 2025). StoryGen (Liu et al., 2024) propagates previous visual-language context during sequential story visualization, while StoryDiffusion (Zhou et al., 2024) introduces consistency-aware attention for long-range image and video generation. Other approaches improve temporal extension or contextual control for longer generation (Kim et al., 2024; Wu et al., 2025; Dalal et al., 2025), while systems such as VideoDirectorGPT (Lin et al., 2023) and HoloCine (Meng et al., 2026) organize multi-scene or multi-shot generation through higher-level narrative structure. Despite these advances, maintaining stable entity identity and scene consistency across independently generated shots, especially with multiple entities or changing backgrounds, remains a challenge.

Synchronized Multimodal Video Production. Multimodal generation has increasingly coupled text, image, video, and audio through cross-modal conditioning and joint generation (Tang et al., 2023; Wu et al., 2023; Kondratyuk et al., 2024). For modalities with explicit temporal structure, some audiovisual methods further improve synchronization between visual events and generated audio (Luo et al., 2023; Liu et al., 2026; Song et al., 2026). Beyond video generation, complete video production requires coordinating heterogeneous modalities across multiple stages, including scripting, visual generation, speech synthesis, and final composition (Chen et al., 2026; Shi et al., 2026b; Wang et al., 2026). Existing automated and agentic systems organize these operations into structured production workflows (Lin et al., 2023; Harry, 2024; Xu et al., 2026b). Across these stages, MoneyPrinterTurbo (Harry, 2024) adjusts generated visuals to match narration duration during final composition, while OpenMontage (Calesthio, 2026) conversely trims or loops narration to planned clip durations, leaving fine-grained cross-modal synchronization underconstrained. Existing approaches largely address synchronization either within a single generative process or during final composition, leaving temporal synchronization among independently generated modalities across production stages less explicitly modeled.

3 Pixelle-Video

As shown in Figure 2, Pixelle-Video is a cross-shot and cross-modal coordination framework for end-to-end video production. The Stateful Storyboard maintains both production specifications and multimodal artifacts, providing a common representation across heterogeneous production stages. Two coordination mechanisms operate through this shared storyboard along distinct axes: *Persistent Visual Memory* coordinates visual context across shots, while *Cross-Modal Temporal Coordination* coordinates across modalities within each shot. Specifically, Persistent Visual Memory constructs reusable subject and setting references and associates them with relevant storyboard shots to provide cross-shot visual constraints. Cross-Modal Temporal Coordination establishes an audio-derived temporal reference for shot-level visual processing and audio-video composition.

3.1 Task Formulation

Given an input X, which may be a topic description, a fixed script, or a collection of user-provided media assets, Pixelle-Video generates a complete video V. We formulate the task as

$$\mathcal{G} : X \to V. \tag{1}$$

Content planning converts the input into an initial storyboard outline containing K temporally ordered shots,

$$\mathcal{S}^{(0)} = \mathcal{P}(X) = \left\{ s_i^{(0)} \right\}_{i=1}^{K}. \tag{2}$$

The production workflow proceeds through four stages: storyboard planning, visual memory construction, multimodal shot execution, and final composition. Storyboard planning initializes the

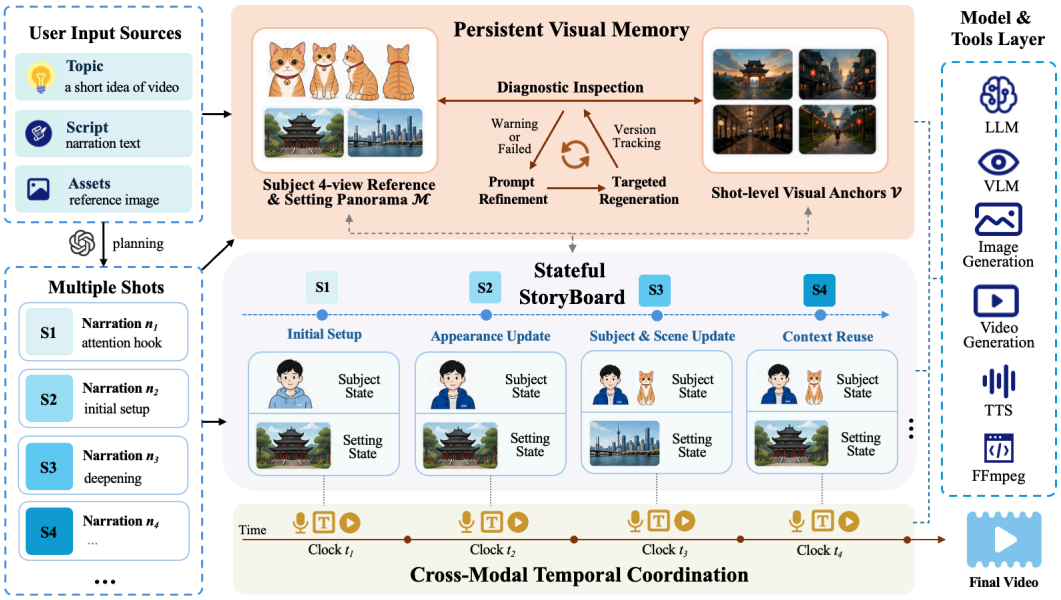


Figure 2: **Overview of Pixelle-Video.** Given a user-provided topic, script, or reference assets, Pixelle-Video organizes the input into a Stateful Storyboard whose shot states are maintained throughout production. *Persistent Visual Memory* establishes reusable cross-shot visual context, while *Cross-Modal Temporal Coordination* propagates audio-derived shot timing across visual realization and composition. Both mechanisms operate through the shared stateful storyboard to coordinate end-to-end video production.

ordered shot structure and its textual content. Visual memory constructs persistent visual references and shot-level visual anchors. Multimodal shot execution realizes each storyboard shot by coordinating its textual content, audio, and visual content under a shared temporal reference. Finally, the resulting clips are assembled into the final video,

$$V = \mathcal{C}\left(c_1, \ldots, c_K; B\right), \tag{3}$$

where c_i denotes the final clip of segment i, and B denoted the optional background music.

3.2 Stateful Storyboard

The initial storyboard $\mathcal{S}$ is progressively materialized into a shared production state as text, audio, temporal, and visual information becomes available throughout the workflow.

We represent the Stateful Storyboard as

$$\mathcal{S} = \left(\Theta, \mathcal{M}, \{s_i\}_{i=1}^{K}\right), \qquad s_i = (n_i, v_i, a_i, t_i, o_i), \tag{4}$$

where Θ denotes task-level production configurations and $\mathcal{M}$ denotes the global visual memory shared across shots. For shot s_i, n_i denotes its textual content; v_i represents the shot-level validated visual anchor; a_i denotes the realized audio; t_i denotes the audio-derived temporal state; and o_i contains the materialized visual, rendered, and composed outputs. We further denote the collection of shot-level visual states as $\mathcal{V} = \{v_i\}_{i=1}^{K}$.

This design exposes two complementary forms of coordination. *Persistent Visual Memory* propagates consistent identity and setting information across storyboard shots through $\mathcal{M}$ and $\mathcal{V}$. *Cross-Modal Temporal Coordination* propagates a shared clock within each shot through t_i to visual generation, rendering, and final composition. The following Sections 3.3 and 3.4 describe these two mechanisms in detail.

3.3 PERSISTENT VISUAL MEMORY

Cross-shot consistency requires recurring subjects and settings to remain recognizable across changes in viewpoint, action, and scene context. Pixelle-Video constructs reusable visual references from the initial storyboard and globally binds them to the shots in which the corresponding subjects and settings appear.

Persistent memory construction. Given the production input and the complete textual content sequence, Pixelle-Video identifies recurring subjects and settings and constructs a persistent visual memory:

$$(\mathcal{E}, \mathcal{L}) = \Phi_{\text{extract}}\left(X,\, \{n_i\}_{i=1}^{K}\right), \qquad \mathcal{M} = \Phi_{\text{mem}}\left(\mathcal{E}, \mathcal{L}\right), \tag{5}$$

where $\mathcal{E}$ and $\mathcal{L}$ denote the recurring subjects and setting inventories, respectively. For each subject, Φ_{mem} constructs a multi-view reference sheet that exposes its visual identity across four dimension viewpoints; for each setting, it constructs a panoramic reference without foreground subjects.

Visual inspection and refinement. Both persistent visual references and shot-level visual anchors are validated through the same diagnostic refinement procedure. Each candidate is inspected against its target specification. Failed candidates are regenerated using diagnostic feedback until passing or reaching the refinement budget. Validated subject and setting references constitute the persistent visual memory $\mathcal{M}$. The same procedure is later applied to shot-level anchor candidates to obtain the validated visual states $\mathcal{V} = \{v_i\}_{i=1}^{K}$.

Global binding. The persistent visual memory is associated with the complete storyboard through a global binding step:

$$\{b_i\}_{i=1}^{K} = \Phi_{\text{bind}}\left(X,\, \mathcal{S},\, \mathcal{M}\right), \tag{6}$$

where b_i denotes the visual binding of shot i to the persistent visual context established for the storyboard. The bindings are determined jointly over the complete storyboard rather than independently for each shot, so that recurring subjects and settings remain associated with consistent visual information across their appearances. The resulting bindings and references are reused during downstream shot realization.

3.4 CROSS-MODAL TEMPORAL COORDINATION

Cross-Modal Temporal Coordination establishes an audio-derived temporal clock for each storyboard shot and propagates it through visual realization, rendering, and audio–video composition. The textual content of each shot is first realized as audio, whose measured duration provides the shared timing basis for subsequent multimodal execution.

Audio-derived temporal clock. For each storyboard shot, the textual content n_i is realized as speech audio with measured duration:

$$(a_i, d_i) = \Phi_{\text{aud}}\left(n_i; \theta_{\text{aud}}\right), \qquad \tau_i = \sum_{j=1}^{i} d_j, \qquad I_i = [\tau_{i-1}, \tau_i), \quad \tau_0 = 0, \tag{7}$$

where a_i denotes the realized audio, d_i its measured duration, and I_i the temporal interval occupied by shot i in the complete video. Together, d_i and I_i define the audio-derived temporal clock t_i for the shot. The clock is written to the Stateful Storyboard and reused throughout subsequent shot execution.

Temporally coordinated shot execution. Given the textual content n_i, realized audio a_i, and validated visual anchor v_i, multimodal shot execution proceeds under the shared temporal clock t_i:

$$o_i = \Phi_{\text{shot}}\left(n_i,\, a_i,\, v_i;\, t_i\right), \tag{8}$$

where Φ_{shot} abstracts visual generation, duration reconciliation, rendering, and audio–video composition. Consequently, the textual, audio, and visual components of each shot are coordinated under the shared temporal clock t_i throughout execution.

Table 1: Quantitative comparison on VBench and ViStoryBench metrics. Higher values indicate better performance for all metrics. Best results among compared methods are highlighted in **bold**, and the second-best results are underlined.

Method	VBench			ViStoryBench			
	Motion Smooth.	Background Consist.	Subject Consist.	Style & Scene Self-Consist.	Individual Action	Character Interaction	Scene Score
MoneyPrinterTurbo	0.9780	0.7308	0.5225	0.4697	1.7012	2.2500	2.4871
OpenMontage	0.9886	0.8598	0.7242	0.5686	2.6424	2.7690	3.1276
ShortGPT	0.9839	0.7874	0.5872	0.4962	2.0388	2.4283	2.7754
Anim-Director	0.9819	0.8517	0.7876	0.7727	2.7582	2.7607	3.1571
ScriptAgent	0.9859	0.8346	0.7004	0.7812	2.3990	2.5064	2.9743
Pixelle-Video	**0.9906**	**0.8927**	**0.8389**	**0.8182**	**3.4340**	**3.5679**	**3.5571**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate end-to-end video production on **DirectorBench** (Chen et al., 2026), a diagnostic benchmark for multi-shot video generation. It covers seven real-world categories: action, cinematic, commercial, educational, narrative, science fiction, and daily-life. Each entry specifies a natural-language instruction together with an optional style descriptor. Due to the cost of full-pipeline generation, we use a subset of 35 instructions (5 per category) as topic-level prompts. All methods receive the same instruction and produce a complete video without human intervention.

Baselines. We compare Pixelle-Video with five representative automated video-production systems, covering both open-source trending GitHub repositories (i.e., MoneyPrinterTurbo (Harry, 2024), OpenMontage (Calesthio, 2026), and ShortGPT (VenturaRay, 2024)) and story-centric generation agents (i.e., ScriptAgent (Mu et al., 2026) and Anim-Director (Li et al., 2024)).

Implementation Details. In our implementation, we set the maximum diagnostic refinement budget to 3 rounds for both persistent visual reference construction and shot-level visual anchor generation. To ensure a fair comparison, we replace each baseline's native LLM, image-generation, video-generation, and speech-synthesis modules of all methods with a shared backbone stack. Specifically, we utilize `deepseek-v4-flash` (Xu et al., 2026a) for scripting and content planning, `qwen3.7-plus` for vision-language modeling, `wan2.7-image` and `wan3.0-video` for image and video generation, and Microsoft Edge TTS for speech synthesis. For all baselines, we follow the official implementations and recommended hyperparameter settings provided in the original papers and open-source repositories.

Evaluation Metrics. To comprehensively evaluate the generated story videos, we utilize three groups of metrics across different benchmarks. For low-level temporal and visual consistency, we use three **VBench** (Huang et al., 2024) metrics: *Motion Smoothness*, *Background Consistency*, and *Subject Consistency* to evaluate temporal quality. For story-level coherence and alignment, we use four **ViStoryBench** (Zhuang et al., 2026) metrics: *Style & Scene Self-Consistency*, *Individual Action*, *Character Interaction*, and *Scene Score*. For cross-modal alignment, we use two **DirectorBench** (Chen et al., 2026) reports *Text–Video Consistency* and *Video–Audio Synchronization* to reflect cross-modal alignment among text, video, and audio. Additional details are provided in Appendix B.

4.2 OVERALL PERFORMANCE

Main Results. Table 1 and Table 2 present the quantitative comparison across VBench, ViStoryBench, and DirectorBench metrics. Pixelle-Video achieves the best performance on all nine evaluated metrics, demonstrating consistent gains across visual consistency, story-level alignment, and cross-modal synchronization. Notably, the improvements are more pronounced on dimensions that require information to remain coherent across shots and modalities, rather than on local clip quality alone. For example, the gain in Motion Smoothness over the strongest baseline is relatively small,

whereas larger improvements are observed in Subject Consistency (0.8389 vs. 0.7876), Character Interaction (3.5679 vs. 2.7690), and Scene Score (3.5571 vs. 3.1571). This pattern is consistent with the role of Persistent Visual Memory, which maintains reusable subject and scene references across storyboard segments and thereby provides persistent visual constraints during generation.

A similar trend appears on DirectorBench metrics, where Pixelle-Video improves Text–Video Consistency from 0.7593 to 0.8197 and Video–Audio Synchronization from 0.5755 to 0.5939 compared with the strongest competing results. These gains are consistent with Cross-Modal Temporal Coordination, which propagates narration-derived timing through visual realization and final composition. Since all systems use the same language, image, video, and speech-generation backbones in our evaluation, the controlled comparison suggests that Pixelle-Video's improvements primarily stem from more effective coordination of visual and temporal information throughout the production process, leading to stronger cross-shot coherence and cross-modal alignment.

Table 2: Quantitative comparison and ablation results on DirectorBench metrics.

Method	Text–Video Consist.	Video–Audio Sync.
MoneyPrinterTurbo	0.3536	0.5694
OpenMontage	<u>0.7593</u>	0.5500
ShortGPT	0.5739	<u>0.5755</u>
Anim-Director	0.6682	0.5510
ScriptAgent	0.6443	0.5643
Pixelle-Video	**0.8197**	**0.5939**
w/o Memory	0.7804	0.5816
w/o Cross-Modal Coord.	0.7125	0.5657

Pairwise Comparison. We further conduct pairwise human preference evaluation between Pixelle-Video and each competing system. As shown in Fig. 3, Pixelle-Video is preferred over every competing system in 70.3%–81.2% of pairwise comparisons, while its loss rate remains below 17.6% across all five baselines. Overall, the consistently high preference rates provide direct perceptual evidence that improved cross-shot and cross-modal coordination translates into higher-quality final videos as judged by human evaluators.

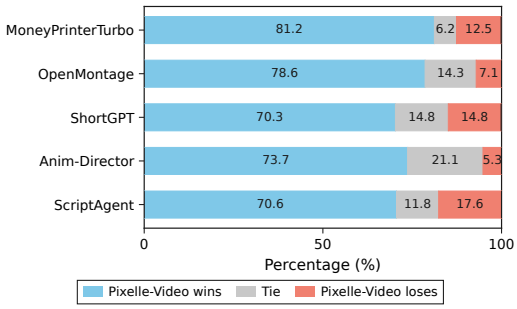


Figure 3: Pairwise human preference between Pixelle-Video and five competing systems. Values denote the percentage of comparisons in which Pixelle-Video wins, ties, or loses.

Human Scoring. To complement the automatic evaluation, we consolidate the nine automatic metrics into six perceptual dimensions for human assessment, covering content fidelity, subject consistency, scene continuity, action and interaction, audio–visual alignment, and motion smoothness, and additionally include overall quality as a holistic measure. Annotation details and scoring guidelines are provided in Appendix C.

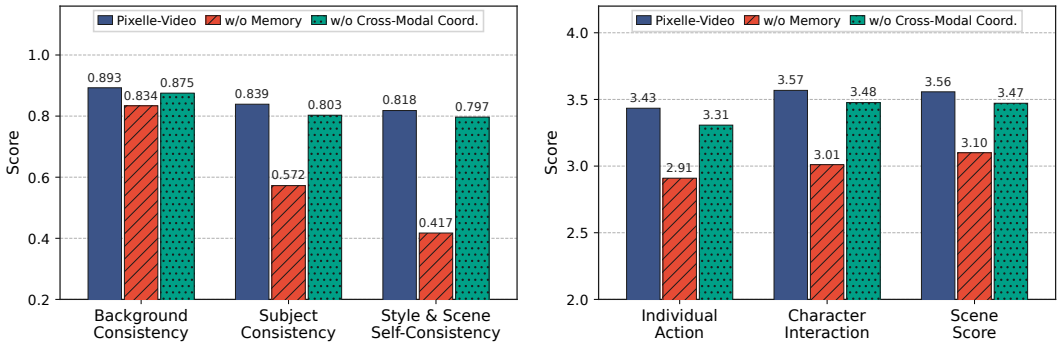
As shown in Table 3, Pixelle-Video achieves the highest score on six dimensions and ties for the highest score on Motion Smoothness. Competing systems tend to excel in specific aspects while showing clear weaknesses in others. For example, OpenMontage performs well in Motion Smoothness (3.4) and Scene Continuity (3.1) but scores only 2.1 in Audio–Visual Alignment. Pixelle-Video instead shows more balanced performance across these dimensions. Its higher Subject Consistency (3.6) and Scene Continuity (3.2) are consistent with Persistent Visual Memory, while its Audio–Visual Alignment score of 3.4 supports the effectiveness of Cross-Modal Temporal Coordination in synchronizing visual content with timing. These improvements lead to an Overall Quality score of 3.2, substantially higher than the strongest competing score of 2.4.

4.3 Ablation Study

Figure 4 summarizes the effects of removing Persistent Visual Memory and Cross-Modal Temporal Coordination. As shown in Fig. 4a, removing visual memory causes the most substantial degradation in visual consistency, reducing Subject Consistency and Style & Scene Self-Consistency by 0.27 and 0.40 points, respectively. In Fig. 4b, Storyboard-level performance also declines substantially, with Individual Action, Character Interaction, and Scene Score decreasing by 0.52, 0.56, and 0.46 points.

Table 3: Human evaluation results across different dimensions. Best results are shown in **bold**, and second-best results are underlined.

Method	Content Fidelity	Subject Consist.	Scene Continuity	Action & Interact.	Audio–Visual Alignment	Motion Smooth.	Overall Quality
MoneyPrinterTurbo	2.1	1.8	1.4	2.8	1.5	2.8	1.8
OpenMontage	2.7	3.3	3.1	3.1	2.1	**3.4**	2.3
ShortGPT	2.2	2.1	2.1	2.9	2.7	3.2	2.1
Anim-Director	2.4	3.4	2.5	2.0	2.5	2.8	2.4
ScriptAgent	2.8	2.9	2.9	2.2	3.1	3.0	2.4
Pixelle-Video	**3.3**	**3.6**	**3.2**	**3.2**	**3.4**	**3.4**	**3.2**



(a) Visual Consistency.

(b) Storyboard Alignment.

Figure 4: Ablation analysis of Pixelle-Video across visual consistency and storyboard alignment.

In comparison, removing Cross-Modal Temporal Coordination causes much smaller changes on these metrics, highlighting the role of persistent visual context in maintaining subjects, scenes, and narrative coherence across shots.

Cross-Modal Temporal Coordination shows a different effect. Table. 2 shows that removing it reduces Text–Video Consistency by 0.11 points and Video–Audio Synchronization by 0.03 points. This confirms the importance of cross-modal coordination for aligning visual realization with narration and temporal structure.

4.4 Case Study

Figure 5 highlights two key aspects of end-to-end video production. In terms of **Content fidelity and cross-modal alignment**, Pixelle-Video consistently preserves the Asian main character, casual-pants product identity, and sales-oriented narrative throughout the sequence, while closely matching the visual content to the corresponding narration. In comparison, OpenMontage remains broadly relevant to the product category, but its narration begins to repeat in later shots, reusing earlier dialogue even as the visual content changes, which leads to clear audio–visual misalignment. ShortGPT does not consistently match the specified Asian male target audience, while Anim-Director adopts a strongly animated visual style that deviates from the requested commercial.

For **cross-shot subject and setting visual consistency**, Pixelle-Video maintains stable subject identity and product appearance across shots while preserving a coherent commercial setting. In contrast, MoneyPrinterTurbo and OpenMontage exhibit substantial subject and setting drift, including changes in model identity, pants appearance, and indoor environments. ScriptAgent maintains a relatively stable model identity, but its first-and-last-frame generation strategy still leads to visual drift over time, including changes in pants color in later shots and noticeable shifts in setting. Additional cases are provided in Appendix A, and the corresponding comparison videos are included in the supplementary material.

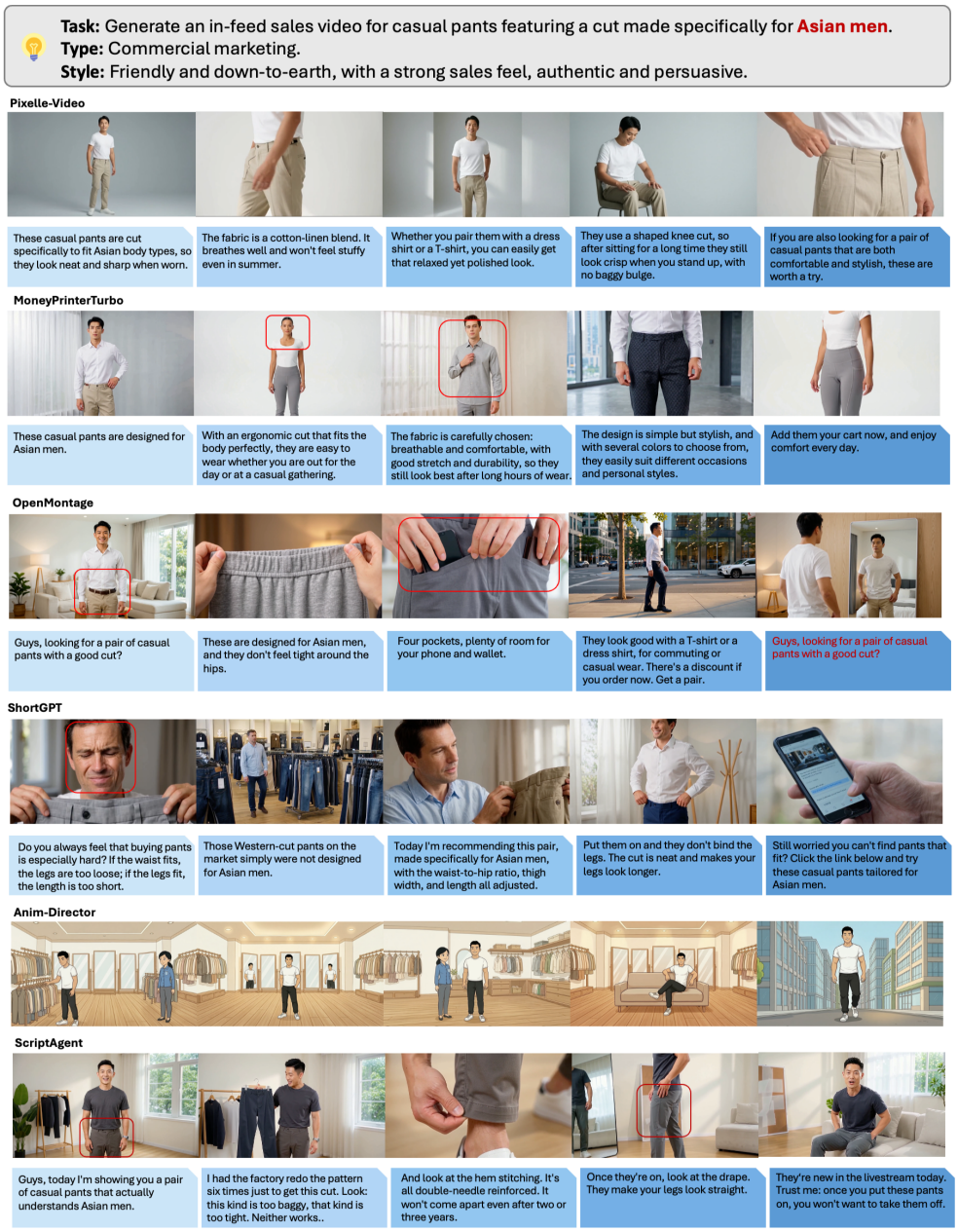


Figure 5: **Comparison on end-to-end commercial video production.** Given the same instruction to generate an in-feed sales video for casual pants designed for Asian men, we compare Pixelle-Video with five baselines. Red boxes highlight representative inconsistencies in competing methods.

5 Conclusion

We introduce Pixelle-Video, a cross-shot and cross-modal coordination framework for end-to-end video production. By combining Persistent Visual Memory for cross-shot consistency with Cross-Modal Temporal Coordination for multimodal alignment, Pixelle-Video enables production stages to share and reuse evolving visual and temporal context. Experiments across VBench, ViStoryBench, and DirectorBench, together with human evaluation and ablation studies, demonstrate consistent improvements in cross-shot coherence and cross-modal alignment. Overall, these results highlight the importance of coordinating visual and temporal state across shots, modalities, and heterogeneous production stages to achieve coherent and high-quality end-to-end video production.

AI USE STATEMENT

In this work, we used generative AI tools during manuscript preparation and revision, including improving grammar, clarity, conciseness, and phrasing of author-written text. Vision-language models were additionally used as automated evaluators in our experimental pipeline, as described in the experimental setup. We did not use generative AI tools for retrieval or discovery of related work, research ideation, or other research-related tasks. The use of generative AI for generating synthetic datasets and proving mathematical claims was not applicable to this work. All AI-assisted edits were reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Pixelle-Video automates multiple stages of video production using generative models, and therefore inherits potential risks associated with generated media. Generated videos may contain inaccurate, biased, or misleading visual and audio content, and should be reviewed before being used in factual or sensitive applications. Human evaluation in this work is limited to assessing the perceptual quality of generated videos under a standardized annotation protocol and does not involve the collection of sensitive personal information.

REPRODUCIBILITY STATEMENT

We describe the task formulation in Section 3.1, and provide detailed formulations of Persistent Visual Memory and Cross-Modal Temporal Coordination in Sections 3.3 and 3.4. Section 4.1 specifies the evaluation dataset, baseline systems and implementation settings. Appendix C provides the human evaluation protocol, scoring rubric, and inter-annotator agreement analysis, while Appendix D includes the prompt templates used for content planning, visual memory construction, shot-level generation, inspection, and regeneration. Additional qualitative examples and the corresponding generated videos are provided in Appendix A and supplementary material.

REFERENCES

Lu Bai, Duo Xu, Qiang Mei, Yongjun Li, Rui Yao, and Peng Wang. Seedance 2.0: Prospects and reflections on the evolution from "tool" to "creator". *The Innovation*, pp. 101564, 2026.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Calesthio. Openmontage. https://github.com/calesthio/OpenMontage, 2026. Accessed: 2026-09-01.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024.

Jiamin Chen, Qianben Chen, Jiawen Zhang, Yidi Wu, Yuchen Li, Xiaokun Zhang, Wangchunshu Zhou, and Chen Ma. Directorbench: Diagnosing long-form video generation with personalized multi-agent evaluation, 2026. URL https://arxiv.org/abs/2605.30090.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17702–17711. IEEE, 2025.

Google DeepMind. Veo 3 technical report. `https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf`, 2025. Technical report.

Harry. Moneyprinterturbo. `https://github.com/harry0703/MoneyPrinterTurbo`, 2024. Accessed: 2026-09-01.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024. doi: 10.1109/CVPR52733.2024.02060.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37: 89834–89868, 2024.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: a large language model for zero-shot video generation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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*, pp. 1–11, 2024.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning, 2023.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6190–6200, 2024.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Scofield Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *International Conference on Learning Representations*, volume 2026, pp. 139160–139194, 2026.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö. Arik. Vista: A test-time self-improving video generation agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6021–6032, June 2026.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36: 48855–48876, 2023.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Shuailei Ma, Yixuan Li, Cheng Chen, et al. Holocine: Holistic generation of cinematic multi-shot long video narratives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 461–471, 2026.

Chenyu Mu, Xin He, Qu Yang, Wanshun Chen, Jiadi Yao, Huang Liu, Zihao Yi, Bo Zhao, Xingyu Chen, Ruotian Ma, et al. The script is all you need: An agentic framework for long-horizon dialogue-to-cinematic video generation. *arXiv preprint arXiv:2601.17737*, 2026.

Haoyuan Shi, Yunxin Li, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. Animaker: Multi-agent animated storytelling with mcts-driven clip generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Haoyuan Shi, Yunxin Li, Nanhao Deng, Zhenran Xu, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. Msvbench: Towards human-level evaluation of multi-shot video generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24034–24058, 2026a.

Yufei Shi, Weilong Yan, Naixuan Huang, Yucheng Chen, Chenyu Zhang, Tao He, Si Yong Yeo, and Ming Li. One sentence, one drama: Personalized short-form drama generation via multi-agent systems, 2026b. URL https://arxiv.org/abs/2605.22144.

Jibin Song, Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Syncphony: Synchronized audio-to-video generation with diffusion transformers. In *International Conference on Learning Representations*, volume 2026, pp. 45373–45398, 2026.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099, 2023.

VenturaRay. Shortgpt. https://github.com/RayVentura/ShortGPT, 2024. Accessed: 2026-09-01.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Shengyi Wang, Niantong Li, Guangzheng Hu, Hong Qi, Fei Ding, Weixu Qiao, Jinlin Wang, Xiaotong Lv, Peng Han, Zimeng Li, Fanshu Ding, Yushu Wang, Han Wu, Jingjing Chen, Chongxiao Wang, Yanhao Wu, Chenglong Huang, Xiaoqian Zhu, Jie Tian, Hua Li, Jingjing Fan, Mingshuang Tang, Zhong Li, Hengxia Qiang, Weibin Chen, Jinyang Zhen, Bing Zhao, Lin Qu, Jing Li, and Hu Wei. Filmbench: A film-grade benchmark for cinematic video generation, 2026. URL https://arxiv.org/abs/2607.24241.

Shuchen Weng, Haojie Zheng, Si Li, Boxin Shi, Xinlong Wang, et al. Audio-sync video generation with multi-stream temporal control. *Advances in Neural Information Processing Systems*, 38: 12845–12865, 2026.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the time: Temporally-controlled multi-event video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23989–24000. IEEE, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Xuenan Xu, Jiahao Mei, Chenliang Li, Yuning Wu, Ming Yan, Shaopeng Lai, Ji Zhang, and Mengyue Wu. Mm-storyagent: Immersive narrated storybook video generation with a multi-agent paradigm across text, image and audio, 2025a. URL https://arxiv.org/abs/2503.05242.

Zhenran Xu, Jifang Wang, Longyue Wang, Zhouyi Li, Senbao Shi, Baotian Hu, and Min Zhang. Filmagent: Automating virtual film production through a multi-agent collaborative framework. In *SIGGRAPH Asia 2024 Technical Communications*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711404. doi: 10.1145/3681758.3698014. URL https://doi.org/10.1145/3681758.3698014.

Zhenran Xu, Xue Yang, Yiyu Wang, Qingli Hu, Zijiao Wu, Longyue Wang, Weihua Luo, Kaifu Zhang, Baotian Hu, and Min Zhang. ComfyUI-copilot: An intelligent assistant for automated workflow development. In Pushkar Mishra, Smaranda Muresan, and Tao Yu (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 632–643, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-253-4. doi: 10.18653/v1/2025.acl-demo.61. URL https://aclanthology.org/2025.acl-demo.61/.

Zhenran Xu, Yiyu Wang, Yunxin Li, Muyang Ye, Xue Yang, Kai Chen, Longyue Wang, Weihua Luo, Baotian Hu, and Min Zhang. ComfyFlow: Benchmarking LLMs for AIGC workflow generation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2903–2916, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.140. URL https://aclanthology.org/2026.findings-acl.140/.

Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Li, Bin Lin, Li Yuan, Lifang He, et al. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*, 2024.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

Zeyu Zhu, Kevin Qinghong Lin, and Mike Zheng Shou. Paper2video: Automatic video generation from scientific papers. *arXiv preprint arXiv:2510.05096*, 2025.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, Xuanyang Zhang, Xianfang Zeng, Zhewei Huang, Gang Yu, and Chi Zhang. Vistorybench: Comprehensive benchmark suite for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9455–9467, June 2026.

APPENDIX

A CASE STUDIES

A.1 CROSS-SHOT SUBJECT CONSISTENCY

Figure 6 illustrates cross-shot subject consistency in a multi-shot educational video centered on an anthropomorphic plane. The instruction explicitly requires the same protagonist to appear throughout the sequence, making identity preservation directly observable. Pixelle-Video maintains a recognizable protagonist across different viewpoints, actions, and settings, preserving its characteristic facial features, body shape, and red-and-white appearance. In contrast, MoneyPrinterTurbo repeatedly switches among realistic gliders, wooden models, and visually distinct aircraft. ShortGPT also exhibits clear subject drift when multiple characters or objects appear, with the intended protagonist being replaced or mixed with other entities across shots. ScriptAgent preserves a similar anthropomorphic plane in the early shots, but gradually introduces visually different aircraft and other different subjects in later scenes. *These examples show that maintaining cross-shot subject identity requires persistent visual constraints beyond local semantic relevance and first-and-last-frame generation.*



Figure 6: **Comparison of cross-shot subject consistency.** Given the same instruction to create an educational video starring an anthropomorphic plane, Pixelle-Video preserves a recognizable protagonist across different viewpoints, actions, and scenes. In comparison, competing systems exhibit varying degrees of subject drift or substitution, including changes in aircraft identity and appearance. Red boxes highlight representative inconsistencies across shots.

A.2 Cross-Shot Scene Consistency

Figure 7 illustrates cross-shot scene consistency in a day-in-the-life vlog progressing from morning to bedtime. Pixelle-Video preserves a recognizable home environment across changing activities, viewpoints, and lighting conditions, retaining recurring cues such as large windows, wooden flooring, green plants, and shelving while allowing natural day-to-night transitions. In contrast, OpenMontage switches between visually unrelated indoor, outdoor, and residential environments. ShortGPT maintains a broadly consistent visual style, but the scenes share limited spatial and environmental continuity across shots. ScriptAgent similarly depicts different residential interiors rather than a persistent home, with noticeable changes in room layout and furniture placement. *These examples highlight that scene consistency requires a shared spatial context across shots despite changes in viewpoint, activity, and lighting.*

B Evaluation Metrics Details

We provide additional details on the nine automatic metrics used in our evaluation. As summarized in Sec. 4.1, the metrics cover three complementary aspects of end-to-end video production: low-level temporal and visual consistency, story-level visual and semantic alignment, and cross-modal alignment. Higher values indicate better performance for all metrics.

For VBench, we directly use the official evaluation protocols. For ViStoryBench, we follow the prior work of (Shi et al., 2026b) and adapt its image-based prompt-alignment evaluation to video by sampling frames at 1 FPS. The sampled frames and the corresponding storyboard description are evaluated by `qwen3.7-plus`. For the DirectorBench cross-modal metrics, we also use `qwen3.7-plus` as the VLM evaluator and follow its checkpoint-based scoring protocol.

VBench. We directly apply the official VBench implementations without modifying their native temporal processing. Motion Smoothness is evaluated using AMT-S frame interpolation, while Background Consistency and Subject Consistency are computed from frame-level visual represen-

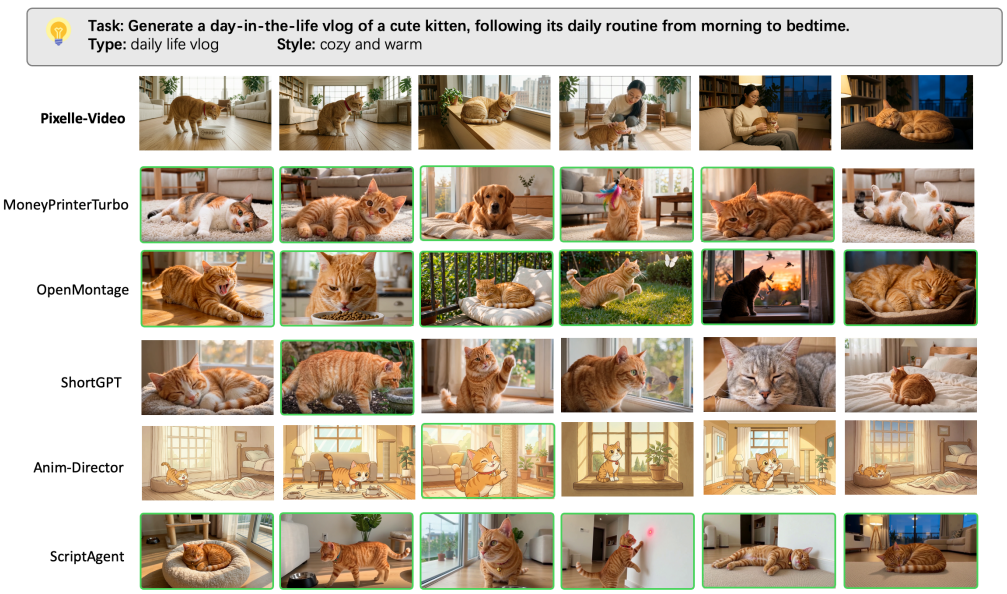


Figure 7: **Comparison of cross-shot scene consistency.** Given the same day-in-the-life vlog instruction, Pixelle-Video preserves a recognizable home environment across changes in activity, viewpoint, illumination, and time of day. Competing systems frequently alter room layout, furniture, architectural cues, or the overall setting across shots. They also often lose fine-grained subject attributes, such as the cat's red collar and gold bell, further weakening cross-shot continuity.

tations using CLIP ViT-B/32 and DINO ViT-B/16, respectively. All three metrics are reported in $[0, 1]$.

ViStoryBench. For the three prompt alignment metrics, we adapt the original image-based ViStoryBench protocol to video following the multi-frame evaluation setting: frames are sampled at 1 FPS and provided, together with the corresponding storyboard description, to `qwen3.7-plus`. Individual Action, Character Interaction, and Scene Score are evaluated on the original five-level 0–4 rubric. For Style & Scene Self-Consistency, we follow the Self-CSD protocol and compute pairwise cosine similarity between CSD ViT-Large style representations, with scores reported in $[0, 1]$.

DirectorBench. We use the Text–Video Consistency and Video–Audio Synchronization metrics for cross-modal evaluation. Following DirectorBench's protocol, `qwen3.7-plus` scores multiple evaluation aspects on a 1–5 ordinal scale, after which the scores are normalized and aggregated into the reported $[0, 1]$ scores.

C Human Evaluation Details

We recruited 15 Ph.D. students in computer science as human annotators. The annotators evaluated the generated videos according to a standardized annotation protocol covering the evaluation dimensions used in our human study. Detailed annotation instructions and scoring guidelines are provided in Table 4.

Before the formal evaluation, all annotators completed a training session to familiarize themselves with the evaluation criteria and rating procedure. We then measured inter-annotator correlations to assess the consistency of their judgments. As shown in Table 5, the annotators exhibit consistently strong agreement across all seven evaluation dimensions. Pearson correlations range from 0.742 to 0.889, Spearman correlations from 0.744 to 0.874, and Kendall's τ from 0.681 to 0.815. The consistently high correlations across the three measures indicate reliable agreement among annotators and support the consistency of our human evaluation protocol.

Table 4: 0–4 Likert rubric used for human evaluation across seven dimensions. Higher scores indicate better performance.

Metric	0	1	2	3	4
Content Fidelity	The requested content is largely absent, contradicted, or replaced by unrelated content.	The video is only weakly related to the instruction, with multiple core requirements missing or incorrect.	The main request is recognizable, but important content is missing, incorrect, or weakly grounded.	Most requested content is correctly realized, with only minor omissions or mismatches.	The requested content is faithfully and unambiguously realized throughout the video.
Subject Consistency	Recurring subjects frequently change identity or appearance and cannot be recognized as the same subject across shots.	Major identity drift occurs across multiple shots, including substantial changes in appearance, clothing, or defining attributes.	The subject remains generally recognizable, but noticeable appearance inconsistencies occur across shots.	Subject identity is preserved across nearly all shots, with only minor variations in non-essential details.	The same subject identity and defining visual attributes are consistently preserved across all relevant shots.
Scene Continuity	Shots depict unrelated or contradictory environments, with no persistent scene identity.	Major changes in layout, furnishings, architecture, or setting repeatedly break scene continuity.	The overall setting is recognizable, but several spatial or environmental details vary noticeably across shots.	The environment remains coherent across shots, with only minor changes that do not disrupt scene identity.	A consistent scene identity is preserved while allowing natural changes in viewpoint, activity, lighting, or time.
Action & Interaction	Required actions or interactions are absent, incorrect, or visually uninterpretable.	Only a small portion of the intended actions or interactions is realized, with major semantic or physical errors.	Core actions or interactions are recognizable, but their realization is incomplete, ambiguous, or partially incorrect.	Most actions and interactions are clearly and correctly realized, with only minor inaccuracies.	Actions, subject relations, and interactions are clearly, naturally, and faithfully realized throughout.
Audio–Visual Alignment	Audio and visual content are largely unrelated or severely temporally misaligned.	Frequent semantic or temporal mismatches occur between the audio and corresponding visual content.	Audio and visuals are broadly related, but noticeable timing or semantic mismatches occur in several segments.	Audio and visual content are well aligned in both semantics and timing, with only minor local mismatches.	Audio and visual content remain consistently synchronized and semantically aligned throughout the video.
Motion Smoothness	Motion is severely unstable, discontinuous, or distorted, substantially impairing viewing.	Frequent jitter, abrupt motion changes, or temporal artifacts noticeably disrupt the video.	Motion is generally understandable, but several visible artifacts or discontinuities remain.	Motion is mostly smooth and natural, with only isolated minor temporal artifacts.	Motion is consistently smooth, stable, and visually natural throughout the video.
Overall Quality	Severe failures in coherence, visual quality, synchronization, or completeness make the video largely unusable.	Multiple major defects substantially impair the overall viewing experience.	The video is generally watchable, but several noticeable issues reduce overall production quality.	The video is coherent and polished overall, with only minor defects that do not substantially affect viewing.	The video is consistently coherent, visually compelling, well synchronized, and of high overall production quality.

Table 5: Inter-annotator correlation across different evaluation dimensions. We report Pearson's r, Spearman's ρ, and Kendall's τ.

Metric	**Pearson** r	**Spearman** ρ	**Kendall** τ
Content Fidelity	0.806	0.760	0.684
Subject Consistency	0.744	0.744	0.681
Scene Continuity	0.822	0.817	0.761
Action & Interaction	0.834	0.819	0.799
Motion Smoothness	0.742	0.745	0.700
Audio–Visual Alignment	0.889	0.874	0.807
Overall Quality	0.840	0.859	0.815

D Prompt Templates

D.1 Content Planning

Content Planning

```
# Role Definition
You are a professional short-video content planner. Given a topic or theme, expand it into
     a sequence of coherent narrations suitable for TTS and subsequent video generation.

# Input
Topic: {topic}
```

```
# Core Task
Generate {n_storyboard} storyboard narrations that develop the topic into a coherent short-
    video narrative.

Each narration should:
- communicate a meaningful viewpoint, explanation, example, or insight;
- sound natural and conversational, like a knowledgeable friend speaking;
- connect naturally with the surrounding narrations;
- provide useful content rather than generic or formulaic statements.

# Narration Requirements
- Language: Match the language of the input topic unless another output language is
    explicitly requested.
- Length: Each narration must contain {min_words}~{max_words} words. For Chinese, count
    Chinese characters rather than whitespace-separated words.
- Purpose: The narration will be converted into speech by TTS, so write for natural spoken
    delivery.
- Tone: Accessible, sincere, engaging, and informative. Avoid academic, stiff, overly
    sentimental, or template-like language.
- Content: Expand around the given topic and ensure each storyboard contributes new
    information or progression.
- Ending: Do not place punctuation at the very end of each narration. Internal punctuation
    should be used naturally for rhythm and pauses.
- References: Relevant authoritative sources or viewpoints may be incorporated when
    appropriate, but do not fabricate sources.
- Do not include URLs, emojis, numeric numbering, or unnecessary filler.

# Storyboard Structure
- The {n_storyboard} narrations should form one continuous narrative rather than
    independent fragments.
- Use a natural progression such as:
attention or context -> key viewpoint -> deeper explanation -> takeaway or inspiration
- This is a semantic guideline, not a fixed sentence template.
- Maintain a consistent speaking voice across all storyboards and ensure natural
    transitions between adjacent narrations.

# Opening Diversity
Choose each opening according to its specific content.
- Avoid fixed opening formulas and repetitive sentence patterns.
- Do not repeatedly begin narrations with the same word or phrase.
- Avoid habitual openings such as "sometimes", "actually", "have you ever", or "imagine".
- Prefer natural openings based on scenes, observations, viewpoints, questions, contrasts,
    or examples.

# Output Requirements
- Output exactly one raw JSON object.
- Do not include Markdown, code fences, explanations, a preamble, or a closing remark.

# Output Format
{
  "narrations": [
    "First narration content",
    "Second narration content",
    "Third narration content"
  ]
}

Before returning the result, verify that:
- the number of narrations equals {n_storyboard};
- every narration satisfies the required length;
- the language matches the input;
- the narrations form a coherent progression;
- openings are not repetitive.
```

D.2 Visual Memory Construction

Subject and Setting Extraction

```
# Role Definition
You are a professional visual asset manager. Given a short-video topic and its shot
    narrations, extract the recurring subjects and physical settings whose visual identity
     must remain consistent across shots.

# Input
```

```
Topic: {topic}
Shot narrations: {narrations_text}

# Core Task
Build a global visual asset baseline for downstream reference generation and shot-level
     visual grounding.

## Subject Extraction
- Extract every individually identifiable subject that recurs across shots and requires a
     consistent appearance.
- Subjects may include people, animals, products, vehicles, devices, or objects.

Rules:
- Do not extract generic groups or crowds.
- Merge different references to the same subject into one canonical entry.
- Do not merge visually distinct subjects.
- If there is no reusable subject, return an empty "subjects" array.

For each subject, provide a static appearance description:
- Person: approximate age, gender presentation, hairstyle/accessories, clothing, body type,
      and stable facial features.
- Animal: species or breed, body proportions, coloration or pattern, and stable accessories
     .
- Product or object: shape, proportions, color scheme, material, and stable markings.

Constraints:
- Describe objective appearance only.
- Do not mention personality, emotion, expression, action, pose, or temporary state.
- Keep the appearance fixed, neat, and intact.
- Do not describe wounds, blood, torn clothing, damage, scars, birthmarks, tattoos, unusual
      face paint, eye color, eyelids, or skin around the eyes.
- Fantasy or science-fiction appearances are allowed when supported by the input.
- Avoid unsupported visual details.
- Use 30-50 words, plain and direct, with no metaphors or literary rhetoric.

## Setting Extraction
Extract every concrete physical setting required by the topic or narrations. For each
     setting, summarize: space type and layout, lighting, core furnishings or structures,
     relevant textures and materials.
- A setting must be a physical place or spatial environment.
- Merge different references to the same location into one canonical entry.
- Keep genuinely different locations separate.
- Do not extract abstract concepts, activities, emotions, or narrative states.
- If no explicit physical setting is given, return an empty "settings" array.

Constraints:
- Describe a stable environment rather than a shot-specific moment.
- Do not include characters, actions, or temporary events.
- Lighting must provide adequate visibility.
- Keep the environment clean, coherent, and visually usable.
- Avoid dirty, bloody, disgusting, or severely dilapidated descriptions.
- Avoid unsupported scene details.
- Use 30-50 words, objective, concrete, and visually shootable.

# Global Consistency Requirements
Before returning the result:
- Consolidate repeated references to the same subject or setting.
- Ensure each subject has one fixed appearance.
- Ensure each setting has one stable spatial layout.
- Remove shot-specific actions, expressions, poses, camera instructions, and temporary
     states.
- Do not create assets unsupported by the input.

# Output Requirements
- Output exactly one raw JSON object.
- Do not include Markdown, code fences, explanations, a preamble, or a closing remark.

# Output Format
{{
  "title": "short video title",
  "logline": "one-sentence story outline",
  "genre": ["genre 1", "genre 2"],
  "mood": "overall mood",
  "subjects": [
    {{
      "name": "canonical subject name",
      "kind": "person or animal or product or object",
      "description": "static appearance description"
    }}
```

```
  ],
  "settings": [
    {{
      "name": "canonical setting name",
      "description": "stable physical setting description"
    }}
  ]
}}
```

Subject Reference Generation Prompt

```
Subject design reference sheet, {style}, pure white background.
Subject name: {name}
Subject type: {kind}
Subject description: {desc}

Generate four views of the same subject in a single row, from left to right:
1) front close-up view,
2) front full view,
3) side full view,
4) back or rear full view.

Adapt the presentation naturally to the subject type:
- For a person, the close-up should show the chest and above, while the remaining views
     show the full body.
- For an animal, the close-up should clearly show the head and upper body, while the
     remaining views show the complete body.
- For a product, vehicle, device, or object, the close-up should clearly show its
     representative front appearance or defining details, while the remaining views show
     the complete object from the specified directions.

Rules:
- Every view must depict exactly the same subject. Keep all identity-defining visual
     features highly consistent across views, including shape, proportions, coloration,
     hairstyle or fur pattern, clothing or accessories, materials, surface finish, and
     stable markings when applicable.
- Keep the subject clean, tidy, intact, and in a neutral static state. Do not show actions,
      dramatic poses, temporary changes, damage, wounds, blood, torn clothing, or broken
     components. For human subjects, do not introduce scars, birthmarks, tattoos, or
     unusual face paint.
- Use a clean and consistent layout with clear separation between the four views.

Absolutely no text, labels, annotations, captions, notes, borders, or watermarks.
```

Setting Reference Generation Prompt

```
Wide panoramic establishing shot, eye-level wide-angle lens, {style}, cinematic composition
     , high quality.
Setting name: {name}
Setting description: {desc}

Generate a clear global reference of the setting that establishes its persistent spatial
     layout, architecture, furnishings, materials, lighting, and other defining
     environmental features.

Rules:
- The main area of the frame must contain no people, animals, or other living beings. Do
     not introduce foreground subjects or character-specific elements.
- Keep the environment clean, tidy, coherent, and intact. It may be plain, modest,
     elaborate, or stylized as required by the description, but it must contain nothing
     dirty, chaotic, disgusting, bloody, severely damaged, or dilapidated.
- Preserve adequate visibility, a complete view of the environment, and a balanced
     composition suitable for reuse as a consistent scene reference across shots.
- Absolutely no text, labels, annotations, captions, notes, or watermarks.
```

Global Binding Prompt

```
# Role Definition
You are a shot-binding assistant. Given the global subject list, setting list, and shot
     narrations, assign the persistent visual assets that should be bound to each shot.
```

```
# Input
Topic: {topic}
Subject list: {subjects_json}
Setting list: {settings_json}
Shot narrations: {narrations_json}

# Core Task
For each shot, determine:
- which listed subjects should appear on screen;
- which listed setting should serve as the physical environment.
Use the topic as global context and the narration of each shot as the primary evidence for
    binding.

# Binding Requirements
1. Output exactly one raw JSON object. Do not include Markdown, explanations, code fences,
    a preamble, or a closing remark.
2. The number of entries in "frames" must exactly equal the number of shot narrations.
3. "index" starts at 0 and increases consecutively in the same order as the narrations.
4. "subjects" may contain only exact "name" values from the provided subject list.
   - Include only subjects that should visibly appear in the shot.
   - Resolve pronouns or alternative references to their canonical subject names when
     supported by the context.
   - If no listed subject should appear, use an empty array.
5. "setting" may contain only an exact "name" value from the provided setting list.
   - Select the setting explicitly or clearly implied by the shot narration and surrounding
     context.
   - If no listed setting can be assigned, use an empty string.
6. Keep the same canonical subject and setting names across all shots.
7. Do not invent, rename, merge, or modify subjects or settings that are not present in the
     provided lists.

# Output Format
Return only the JSON object.
{{
  "frames": [
    {{
      "index": 0,
      "subjects": ["subject name"],
      "setting": "setting name"
    }}
  ]
}}
```

Shot-Level Visual Anchor Generation Prompt

```
# Role Definition
You are a professional AI prompt engineer and film storyboard director. Given one shot, the
     relevant subject descriptions, and the setting description, write an English prompt
    for generating the static first frame of the shot.

# Core Task
Create a high-quality still image that establishes the initial visual state before any
    motion occurs.

Use the shot narration as context, but convert dynamic events into a stable composition.
    Clearly describe:
- visible subjects and their positions;
- stable pose, body orientation, gaze, and shot-relevant expression;
- spatial relations among subjects and nearby objects;
- setting, lighting, camera framing, and composition.

# Static-Frame Requirements
- Describe only a still image.
- Do not describe actions in progress, motion paths, or the onset of an action.
- Convert actions into stable states, e.g., "standing beside the door", "sitting on a chair
    ", or "hands resting on the table".
- Do not use phrases such as "about to", "preparing to", "just about to", or "in the middle
     of".

# Subject and Setting Grounding
- Preserve the persistent appearance of all provided subjects. Do not change identity-
    defining clothing, hairstyle, body features, colors, materials, or markings.
- Do not invent new subjects.
- Preserve the provided setting and its main spatial structure.
- Keep subject placement physically plausible and avoid clipping or impossible overlap with
     tables, chairs, doors, walls, vehicles, or props.
```

```
- In multi-subject shots, make relative positions and gaze relations clear.

# Camera and Visual Quality
- Specify the shot scale, camera position or angle, composition, and placement of the main
    subject or subjects.
- Use cinematic composition, coherent lighting, natural spatial depth, clear subject
    hierarchy, and stable high-quality detail.

# Prohibited Content
- No actions in progress or future actions.
- No hidden intentions or internal thoughts.
- No changes to persistent subject or setting identity.
- No text, subtitles, labels, UI elements, or watermarks.
- No explanations or Markdown.

# Input
Full overview: {original_text}
Current shot narration: {plot}
Subject descriptions: {subject_description}
Setting description: {setting_description}

# Example
- Negative: A man is running toward the window, about to jump out, shouting at the street
    outside.
- Positive: Cinematic medium shot. A man stands in front of a large window, body angled
    slightly toward the glass, gaze directed at the street outside. Both feet are planted,
     arms resting naturally at his sides, with a clear gap between him and the window
    frame. Natural directional light, balanced composition, stable high-detail still image
    .

# Output Requirements
- Output only the final English visual prompt.
- Do not include explanations, Markdown, or additional text.
```

Subject Reference Inspection Prompt

```
# Role Definition
You are a professional evaluator of AI-generated subject reference images. Judge whether
    the image can serve as a stable subject asset for later video generation.

Subject description: {subject_description}

# Expected Layout
The image should contain four consistent views of the same subject:
1. front close-up or defining-detail view;
2. front full view;
3. side full view;
4. back or rear full view.

# Hard Failures
If any hard failure is present, "is_acceptable" must be false.
- Any required view is missing or clearly incomplete.
- The views do not depict the same subject or show major identity changes.
- Core attributes conflict with the subject description, such as age or gender for a person
    , species for an animal, or shape, color, material, clothing, hairstyle, or markings
    when applicable.
- The subject is severely deformed, broken, occluded, or cropped and cannot serve as a
    reference.
- Text, labels, annotations, captions, watermarks, or UI elements are visible.
- The image is severely blurry, overexposed, too dark, or otherwise unreadable.

# Soft Evaluation
If there is no hard failure, evaluate:
- fidelity to the subject description;
- consistency of identity-defining details across views;
- clarity of the front, side, and rear views;
- readable and unobstructed composition;
- consistency with the intended visual style;
- overall usefulness as a reusable subject reference.
Minor flaws must remain soft issues and must not be treated as hard failures.

# Acceptance
- If "hard_failures" is non-empty: "is_acceptable" = false and "score" <= 6.
- If "hard_failures" is empty and "score" >= 7: "is_acceptable" = true.
- Otherwise: "is_acceptable" = false.
```

```
# Output Format
Return exactly one raw JSON object:
{
  "score": 1-10,
  "hard_failures": ["hard failure"],
  "soft_issues": ["soft issue"],
  "issues": ["main problem"],
  "is_acceptable": true/false,
  "suggestion": "What should be fixed first if regeneration is needed; otherwise an empty
     string"
}
```

Setting Reference Evaluation Prompt

```
# Role Definition
You are a professional evaluator of AI-generated setting reference images. Judge whether
     the image can serve as a stable environment asset for later video generation.

Setting description: {setting_description}

# Hard Failures
If any hard failure is present, "is_acceptable" must be false.
- The image depicts the wrong type of setting.
- The main spatial structure is unreadable or unusable as an environment reference.
- A defining object, structure, or spatial element is missing or clearly incorrect.
- Unrelated people, animals, or other prominent living subjects appear. If an empty setting
      is required, any visible person or animal is a hard failure.
- Text, labels, annotations, captions, watermarks, or UI elements are visible.
- The image is severely blurry, overexposed, too dark, distorted, or has unusable
     perspective.

# Soft Evaluation
If there is no hard failure, evaluate:
- layout and object-placement fidelity;
- quality and clarity of key environmental elements;
- lighting and atmosphere;
- color and style consistency;
- spatial perspective and depth;
- composition and reuse value;
- sufficient environmental detail without excessive clutter.
Minor layout differences, ordinary lighting, or slightly sparse detail are soft issues only
     .

# Acceptance
- If "hard_failures" is non-empty: "is_acceptable" = false and "score" <= 6.
- If "hard_failures" is empty and "score" >= 7: "is_acceptable" = true.
- Otherwise: "is_acceptable" = false.

# Output Format
Return exactly one raw JSON object:
{
  "score": 1-10,
  "hard_failures": ["hard failure"],
  "soft_issues": ["soft issue"],
  "issues": ["main problem"],
  "is_acceptable": true/false,
  "suggestion": "What should be fixed first if regeneration is needed; otherwise an empty
     string"
}
```

Shot-Level Visual Anchor Evaluation Prompt

```
# Role Definition
You are a professional evaluator of video first-frame references. Judge whether the
     generated image is suitable as the visual anchor and first-frame input for a video
     model.
Shot narration: {plot}
Shot visual prompt:{visual_prompt}
Subject descriptions:{subject_description}
Setting description:{setting_description}

# Hard Failures
If any hard failure is present, "is_acceptable" must be false.
```

```
- A required main subject is missing, or an unrelated main subject is added.
- A main subject is severely broken, unrecognizable, occluded, or improperly cropped.
- A subject's core identity conflicts with its reference description.
- The setting is clearly inconsistent with the required environment.
- Subject pose, placement, gaze, or core relations are fundamentally inconsistent with the
    intended first-frame state.
- Shot scale or composition is unusable for video generation.
- A subject visibly clips through, intersects with, floats above, or makes implausible
    contact with furniture, walls, doors, vehicles, props, or other scene elements.
- Text, labels, annotations, captions, watermarks, or UI elements are visible.
- The image is severely blurry, overexposed, too dark, distorted, or otherwise unusable.

# Soft Evaluation
If there is no hard failure, evaluate:
- clarity of the static subject state;
- subject placement, orientation, gaze, and interaction;
- consistency with subject references;
- consistency with the setting reference;
- camera scale, angle, and composition;
- physical plausibility of subject-setting relations;
- lighting and visual atmosphere;
- overall image quality and usefulness for image-to-video generation.
Do not fail an image only for minor lighting differences, sparse background detail, or
    failure to depict dialogue literally.

# Acceptance
- If "hard_failures" is non-empty: "is_acceptable" = false and "score" <= 6.
- If "hard_failures" is empty and "score" >= 7: "is_acceptable" = true.
- Otherwise: "is_acceptable" = false.

# Output Format
Return exactly one raw JSON object:
{
  "score": 1-10,
  "hard_failures": ["hard failure"],
  "soft_issues": ["soft issue"],
  "issues": ["main problem"],
  "is_acceptable": true/false,
  "suggestion": "What should be fixed first if regeneration is needed; otherwise an empty
    string"
}
```

Regeneration Template

```
{base_prompt}
The previous generation did not pass evaluation. In the next generation, prioritize
    correcting the issues below while preserving the original subject identities, setting,
     and shot specification.

Hard failures: {hard_failures}
Main issues: {issues}
Soft issues: {soft_issues}
Revision suggestion: {suggestion}

Regenerate the image with the identified problems corrected. Do not introduce unrelated
    changes to visual attributes that were not flagged by the evaluation.
```