DRAMADIRECTOR: GEOMETRY-GUIDED SHORT DRAMA GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Short dramas combine rapid shot rhythms, dialogue-driven focus shifts, and frame-level spatial constraints that are often underspecified by text-only generation pipelines. We study plot-to-short-drama generation, where a global plot and local context are transformed into a structured storyboard and then realized as a multi-shot video. We propose **DramaDirector**, a geometry-guided framework that connects storyboard planning to visual realization through two coupled components. First, it retrieves appearance-reduced depth–pose references from real short-drama shots to provide reusable priors for spatial layout and body configuration. Second, it uses the learned retrievability of generated storyboard conditions as a GRPO training signal, encouraging the planner to produce descriptions that can be grounded by the downstream visual pipeline. Static visual and dynamic narrative conditions serve as the interface for first-frame generation and image-to-video synthesis. We also introduce **DramaBoard**, a benchmark built from 35 live-action dramas and 2,826 episodes with shot annotations (101,763 shot records); its Continuation subset contains 2,807 episodes and 81,283 shots, with structured storyboards and multi-dimensional evaluation protocols. In the reported benchmark tables, DramaDirector has higher or competitive values than representative agentic and video-generation baselines on several narrative-quality, instruction-following, and intrinsic-video-quality metrics. Our code is anonymously released at: `https://anonymous.4open.science/r/DramaDirector-7484/`

1 INTRODUCTION

Short dramas have become a prominent form of online storytelling, marked by mobile-native distribution, fast pacing, and emotional cliffhangers (Cao et al., 2026). As cinematic video generation advances toward short-movie synthesis with coherent keyframes and cinematic rhythm (Xiao et al., 2025b; Huang et al., 2025a), LLMs serve as a flexible interface for turning plots into structured narrative plans, making them a natural fit for storyboard-level short-drama generation (Lei et al., 2025; Teleki et al., 2025). We focus on *plot-to-short-drama generation*, which converts a global plot and local context into a multi-shot storyboard that is rendered into a short-drama video.

Recent narrative-video systems improve planning through script-level supervision (Tang et al., 2024), decompose production decisions across multiple agents (Wu et al., 2025; Huang et al., 2026), or condition synthesis on structured visual elements (Fei et al., 2025). These approaches make story planning more explicit, but the planner still commonly communicates with the image or video generator through text. As a result, the storyboard may describe what a shot should depict without providing an executable representation of how subjects should be arranged in the frame.

In a common deployment pipeline, an LLM drafts a storyboard, a text-to-image model renders each shot's first frame, and an image-to-video model animates it, exposing two related challenges.

A. Generic storyboards may not follow short-drama shot grammar. Without task-specific supervision, LLM-generated storyboards can become novelistic or underspecified (Zhang et al., 2025a; An et al., 2025), omitting production patterns such as fast cuts, over-the-shoulder dialogue, stacked close-ups, and one-beat-per-shot emotional progression.

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.

B. Text storyboards underspecify frame-level visual structure. Shot descriptions can name subjects, actions, and camera attributes while leaving their screen-space positions, relative depth, body configurations, and occlusion cues ambiguous. These ambiguities are difficult to resolve through text-based planning alone, even when the downstream generator is optimized with text–video alignment objectives (Prabhudesai et al., 2025; Li et al., 2025b).

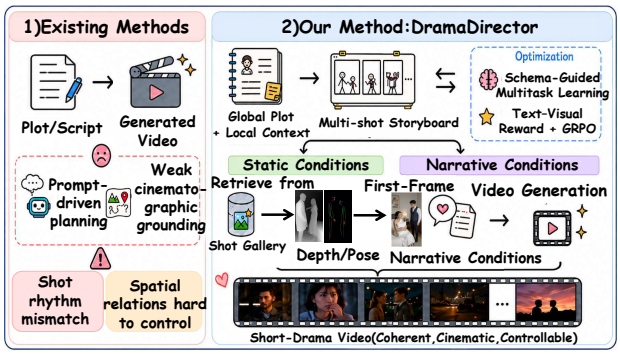

Figure 1: Planning via depth–pose retrieval.

To address the missing interface between planning and visual realization, we present DramaDirector, a geometry-guided framework for real-shot structural retrieval. A trainable alignment model matches each static storyboard condition to a gallery of depth maps and pose skeletons extracted from short-drama shots. These modalities reduce appearance and identity cues while retaining spatial layout and body configuration, so the match is structural rather than semantic. The retrieved reference guides first-frame generation, while the dynamic narrative condition specifies the action, dialogue, and emotion. The same alignment score is used as a GRPO reward during planner training. This closes the loop between planning and generation: the planner is optimized not only for schema-valid text, but also for storyboard conditions that the downstream pipeline can ground in visual evidence.

To validate plot-to-short-drama generation, benchmarks must cover not only final video quality, but also shot-level planning and storyboard-to-video controllability. We therefore introduce DramaBoard, a benchmark constructed from 35 live-action dramas with 2,826 episodes containing shot annotations and 101,763 shot records. The Continuation subset used for the reported experiments contains 2,807 episodes and 81,283 segmented shots. Through episode-level preprocessing, transcript correction, and shot-level storyboard annotation, DramaBoard provides aligned keyframes, dialogue, episode metadata, structured storyboards, and evaluation protocols for storyboard narrative quality, storyboard-to-video instruction following, and intrinsic video quality.

Our contributions are summarized as follows:

- We introduce an appearance-reduced structural retrieval interface that matches storyboard conditions to depth–pose references drawn from real short-drama shots and uses those retrieved references as structural priors for both first-frame synthesis and subsequent video generation.
- We couple the planner to this interface through a learned visual-retrievability reward, using multi-task storyboard supervision and GRPO to optimize storyboard conditions so the downstream retrieval and generation pipeline can ground them.
- We introduce DramaBoard, a benchmark with aligned keyframes, dialogue, structured shot-level storyboards, and protocols for evaluating narrative planning, instruction following, retrieval, and video quality, where aesthetics rises 14.3% over the strongest baseline.

2 Related Work

Short-Drama Generation. Short-drama generation maps compact narrative inputs to dense shot sequences, shooting scripts, and dialogue-centered visual scenes. Recent work explores task-relevant formatting schemes and controllable generation frameworks for this setting (Zhang et al., 2025b; Wang et al., 2026). SkyScript studies script–shooting-script structure for short dramas (Tang et al., 2024), while SkyReels-A2 conditions video synthesis on multiple structured visual elements (Fei et al., 2025). These methods motivate explicit planning and control, but they lack the same planner-to-generator interface as DramaDirector: our planner produces static and dynamic conditions, grounding the static condition through a depth–pose gallery encoding reusable frame-level structure.

RL Post-training. Recent foundation-model research uses reinforcement learning to align generation behavior, with video-specific RL improving temporal reasoning and GRPO-style refinement advancing multimodal alignment (Li et al., 2025a; DeepSeek-AI et al., 2025; Huang et al., 2025b). In visual

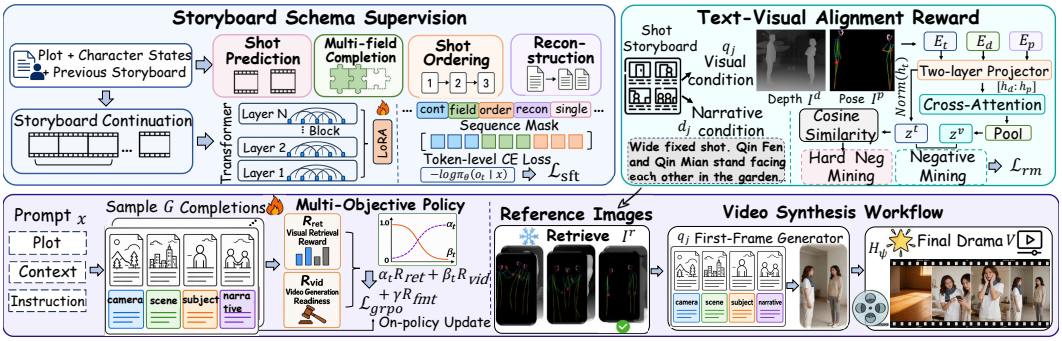

Figure 2: Overall architecture of the proposed DramaDirector framework for short drama generation.

generation, Gen-Searcher, VisionCreator, and VisionCreator-R1 apply GRPO to multi-step visual creation (Feng et al., 2026; Lai et al., 2026a;b). DramaDirector differs in the post-training signal: its policy reward measures whether a generated static storyboard condition retrieves compatible short-drama geometry, rather than generic task or image-quality feedback.

Multimodal RAG. Multi-RAG integrates video, audio, and text streams for adaptive video understanding, while structured RAG frameworks capture long-range spatio-temporal dependencies in video evidence (Arefeen et al., 2024; Mao et al., 2025; Fu et al., 2026). Recent work further analyzes retrieval configurations, reranking, generation-time evidence integration, and retrieval over interleaved text-image documents aligned with downstream VLM preferences (Hu et al., 2025). DramaDirector uses retrieval for a different purpose: it does not treat generic documents, RGB images, or motion examples as the primary control signal, but retrieves depth–pose references that encode short-drama spatial layout and character configuration for first-frame generation. The resulting retrieval score is also reused as a policy-level training signal, linking evidence selection to storyboard planning.

3 Methodology

This section introduces our DramaDirector framework, whose architecture is depicted in Figure 2.

3.1 Geometry-Guided Drama Generation

Given a global plot summary $\mathbf{p} \in \mathcal{P}$ and a local narrative context $\mathbf{c} \in \mathcal{C}$, short-drama generation produces a multi-shot video $\mathcal{V} = \{V_j\}_{j=1}^{N}$. Text carries the narrative, while spatial layout, relative depth, body configuration, and screen-space position come from visual cues. The interface is a structured storyboard $\mathcal{Y} = \{(q_j, d_j)\}_{j=1}^{N}$: the static condition q_j packs camera fields, scene background, and keyframe subject state (screen position, clothing, action, and expression) for retrieval and first-frame generation, excluding character names, dialogue, speaker, emotion, and duration; the dynamic condition d_j packs the shot narrative, dialogue, speaker, and emotion as the video instruction. Because textual q_j leaves the frame configuration ambiguous, the planner g_ϕ maps $(\mathbf{p}, \mathbf{c})$ to $\mathcal{Y}$ and retrieves a depth–pose reference I_j^r from a pre-indexed gallery $\mathcal{G}$ of real short-drama shots (Appendix A.5) as a structural prior for keyframe I_j^0:

$$g_\phi : (\mathbf{p}, \mathbf{c}) \mapsto \mathcal{Y}, \quad I_j^r = \arg\max_{I_i \in \mathcal{G}} (\mathbf{z}_j^t)^\top \mathbf{z}_i^v, \quad I_j^0 = F_\xi(q_j, I_j^r). \tag{1}$$

where $\mathbf{z}_j^t$ and $\mathbf{z}_i^v$ are text- and visual-side embeddings whose inner product measures compatibility in the learned depth–pose alignment space. The first-frame generator F_ξ conditions on q_j and uses I_j^r to provide priors for layout, body configuration, and relative depth; the reference guides these properties without guaranteeing exact geometric preservation.

With the keyframe, an image-to-video model H_ψ rolls out the shot using I_j^0 as the visual anchor and d_j as the temporal instruction, and the final drama is obtained by concatenating all shots:

$$V_j = H_\psi(I_j^0, d_j), \quad \mathcal{V} = (V_1, \ldots, V_N), \tag{2}$$

3.2 STORYBOARD SUPERVISED FINETUNING

The framework above hinges on the storyboard planner g_ϕ, yet off-the-shelf LLMs are not explicitly trained for the camera grammar, field structure, and shot-to-shot transitions of short-drama storyboards. We therefore adapt g_ϕ via supervised finetuning on real short-drama storyboards.

3.2.1 MULTI-TASK STORYBOARD SUPERVISION

Since no single training signal can capture all aspects of storyboard quality, we decompose our SFT task into one main objective and four auxiliary objectives spanning local, structural, and temporal competence, $\mathcal{T}_{\text{sft}} = \{\tau_{\text{cont}}, \tau_{\text{single}}, \tau_{\text{field}}, \tau_{\text{order}}, \tau_{\text{recon}}\}$, each objective detailed as below.

Each episode $s = (\mathbf{p}, \mathbf{c}, \mathcal{Y})$ in the structured storyboard corpus $\mathcal{D}_0$ ($\mathcal{Y} = \{(q_j, d_j)\}_{j=1}^{N}$) is built by automatic shot segmentation, keyframe–dialogue alignment, and schema-conditioned VLM annotation. Task renderer $\mathcal{A}_\tau$ yields its prompt–target pair:

$$(x_\tau, \mathcal{Y}_\tau) = \mathcal{A}_\tau(s), \quad \mathcal{D}_\tau = \{(x_\tau, \mathcal{Y}_\tau)\}_{s \in \mathcal{D}_0}, \tag{3}$$

Storyboard Continuation. Given $\mathbf{p}$, $\mathbf{c}$, and past shots $q_{<j}$, the main task τ_{cont} autoregressively predicts the next w storyboard entries $\{(q_\ell, d_\ell)\}_{\ell=j}^{j+w}$, mirroring inference-time use of g_ϕ.

Auxiliary Objectives. (i) **Single-shot Prediction** predicts the next shot from local context to improve local realization; (ii) **Multi-field Completion** recovers missing fields from partial specifications to capture cross-field dependencies; (iii) **Shot Ordering** reorders shuffled sequences to strengthen temporal coherence; and (iv) **Summary-guided Reconstruction** reconstructs storyboards from summaries to reinforce plan-to-shot grounding. These tasks stably initialize reinforcement learning.

3.2.2 TRAINING OBJECTIVE

We optimize the base model with LoRA adapters on the attention and feed-forward projections. For each training sample, let $\mathcal{O} = (o_1, \ldots, o_n)$ be the tokenized target storyboard and m_t indicate response tokens after the prompt boundary. The objective can be formalized as follows:

$$\mathcal{L}_{\text{sft}} = -\sum_{t=1}^{n} m_t \log \pi_{\theta_{\text{llm}}}(o_t \mid x, o_{<t}), \tag{4}$$

where x is the task prompt and $o_{<t}$ denotes previous target tokens. For short auxiliary tasks, multiple samples are packed into one training sequence under the maximum length constraint, with attention restricted within each sample's boundary so that the loss is computed independently per sample.

3.3 RL FROM RETRIEVAL AND RENDERING

To align g_ϕ with retrieval from $\mathcal{G}$ and rendering by H_ψ, we continue training the SFT-initialized planner with reinforcement learning. We adopt Group Relative Policy Optimization (GRPO), whose value-network-free advantage estimation fits the long, structured nature of storyboard outputs.

Visual Retrieval Reward. For generated shots $\{y_j\}$ paired with target episode e, each static condition q_j is mapped to its text-side alignment vector $\mathbf{z}_j^t$ by the alignment model introduced in Section 3.4. Let $\mathcal{Z}_e^{\pm}$ denote the visual-vector sets from the target and sampled negative episodes. We compute top-k retrieval similarities on each side:

$$\mathcal{S}_j^{\pm} = \text{TopK}\{(\mathbf{z}_j^t)^\top \mathbf{z}^v : \mathbf{z}^v \in \mathcal{Z}_e^{\pm}\}, \tag{5}$$

where TopK returns the k largest scalar similarities. In the released training script, $k = 1$; the positive set is all cached depth–pose vectors from the target episode, while one negative episode is sampled uniformly from another episode of the same drama, otherwise from an episode of another drama. Shot- and completion-level scores are then aggregated into a contrastive reward:

$$R_{\text{ret}} = \tfrac{1}{2} + \tfrac{\bar{c}^+ - \bar{c}^-}{4}, \;\; c_j^{\pm} = \tfrac{1}{k} \sum_{a \in \mathcal{S}_j^{\pm}} a, \;\; \bar{c}^{\pm} = \tfrac{1}{N} \sum_{j=1}^{N} c_j^{\pm}. \tag{6}$$

This contrastive form is intended to reward conditions that match the target episode more strongly than sampled unrelated episodes, thereby discouraging descriptions that retrieve broadly across episodes. Because the reward is a relative (positive-versus-negative) score rather than an absolute similarity maximizer, and because it is combined with the format and video-readiness rewards on top of the SFT-initialized planner, it does not optimize a single similarity scalar in isolation.

Video-Generation Readiness Reward. Beyond static grounding, video synthesis requires dynamic conditions that are concrete and temporally coherent. We evaluate the sequence $\{d_j\}_{j=1}^{N}$ with an LLM judge J, formally defined as:

$$R_{\text{vid}} = J(\{d_j\}_{j=1}^{N}) \in [0, 1], \tag{7}$$

where J scores action concreteness, visual detail, transition coherence, emotional progression, and short-drama pacing over the full sequence.

Reward Aggregation. We first compute a field-coverage format score $R_{\text{fmt}} \in [0, 1]$ over the parsed storyboard, penalizing missing keys, empty content fields, and incomplete subject subfields, and map it to the signed term $\widetilde{R}_{\text{fmt}} = 2R_{\text{fmt}} - 1$ used by the released implementation. It is not a full enum, range, or semantic validator. The three terms are combined as:

$$R(\mathcal{Y}) = \alpha_s R_{\text{ret}} + \beta_s R_{\text{vid}} + \gamma \widetilde{R}_{\text{fmt}}, \;\; \alpha_s = \alpha_0 + (\alpha_{\text{end}} - \alpha_0)p_s, \;\; \beta_s = \beta_0 + (\beta_{\text{end}} - \beta_0)p_s. \tag{8}$$

where s is the current update step, $p_s = \min(s/H_{\text{ann}}, 1)$, H_{ann} is the annealing horizon, and $\alpha_0, \alpha_{\text{end}}$ ($\beta_0, \beta_{\text{end}}$) are the start/end weights of linearly annealed α_s (β_s); γ stays fixed.

Group Relative Policy Optimization. Given a prompt x, we sample a group of G candidate completions and then optimize the clipped objective as follows:

$$\mathcal{L}_{\text{rl}} = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{n_i}\sum_{t=1}^{n_i}\min(r_{i,t}A_i, \bar{r}_{i,t}A_i), \tag{9}$$

where n_i is the length of each response and A_i is obtained by normalizing $R(\mathcal{Y}_i)$ across the G sampled completions. The ratio $r_{i,t}$ compares the current policy with the sampling policy at token t, and $\bar{r}_{i,t}$ clips it to the interval $[1 - \epsilon, 1 + \epsilon]$ with clipping coefficient ϵ.

3.4 Geometry-Aware Text-Visual Alignment

Geometry-guided retrieval (Section 3.1) and the retrieval reward (Section 3.3) both score how well a static condition matches a candidate shot's frame-level structure. RGB vision-language encoders can entangle appearance with geometry and favor semantically similar shots with different layouts, so we train a dedicated alignment model on static conditions and depth–pose evidence, frozen as the shared backbone for inference-time retrieval and policy-time reward.

Textual Condition Decomposition. Each storyboard includes a static visual condition and a dynamic narrative condition. The static condition is $q_j = [a_j; e_j; u_j]$, where a_j is the camera, e_j the scene, and u_j the keyframe subject state (screen position, clothing, action, and expression). It is the prompt for first-frame generation. The dynamic condition d_j collects the narrative description, dialogue, speaker, and emotion, and drives video synthesis. Only q_j participates in alignment, since geometric grounding is a property of the keyframe rather than the temporal rollout.

Architecture. For shot y_j, the visual evidence $I_j = (I_j^d, I_j^p)$ consists of the depth map and the human-pose map of the keyframe. Compared with raw RGB, these modalities reduce appearance and identity cues while retaining information about relative depth, spatial layout, and body configuration. Encoders E_t, E_d, and E_p embed the inputs in $\mathbb{R}^D$, and modality-specific two-layer projection blocks (normalization, GELU, and dropout) map them to D_a dimensions:

$$(\mathbf{x}_j^t, \mathbf{x}_j^d, \mathbf{x}_j^p) = (E_t(q_j), E_d(I_j^d), E_p(I_j^p)), \;\; \mathbf{h}_j^r = P_r(\mathbf{x}_j^r), \; r \in \{t, d, p\}. \tag{10}$$

The text projection is ℓ_2-normalized to form the text-side vector, while the two visual projections are stacked as modality tokens and fused by the self-attention mechanism:

$$\mathbf{z}_j^t = \text{norm}(\mathbf{h}_j^t), \;\; \mathbf{M}_j = [\mathbf{h}_j^d; \mathbf{h}_j^p], \;\; \mathbf{z}_j^v = \text{norm}\big(W_v\,\text{Pool}(\text{LN}(\mathbf{M}_j + \text{MHA}(\mathbf{M}_j))) + \mathbf{b}_v\big). \tag{11}$$

where Pool averages the two modality tokens, and $W_v, \mathbf{b}_v$ are the visual output projection parameters. The resulting $\mathbf{z}_j^t$ and $\mathbf{z}_j^v$ are exactly the vectors consumed by the gallery retriever in Section 3.1 and by the retrieval reward R_{ret} in Section 3.3.

Contrastive Learning. We train the alignment model with bidirectional contrastive learning. Positive pairs are the text and visual vectors of the same shot; negatives are drawn from non-adjacent shots, different episodes, and different dramas to cover progressively coarser mismatches. Given anchors $\{\mathbf{a}_i\}_{i=1}^B$, positives $\{\mathbf{b}_i\}_{i=1}^B$, and negative sets $\{\mathcal{N}_i\}_{i=1}^B$, the contrastive learning is defined as:

$$\kappa_i(\mathbf{b}) = \exp(\cos(\mathbf{a}_i, \mathbf{b})/\tau_{\text{nce}}), \;\; \ell(\mathbf{a}, \mathbf{b}, \mathcal{N}) = -\tfrac{1}{B}\sum_{i=1}^{B} \log \frac{\kappa_i(\mathbf{b}_i)}{\sum_{\mathbf{c}\in\{\mathbf{b}_i\}\cup\mathcal{N}_i} \kappa_i(\mathbf{c})}. \tag{12}$$

where B is the batch size and τ_{nce} the temperature. Letting $\mathcal{N}_i^v$ and $\mathcal{N}_i^t$ denote K sampled visual and textual negatives for anchor i, the bidirectional InfoNCE loss is defined as follows:

$$\mathcal{L}_{\text{nce}} = \tfrac{1}{2}\big[\ell(\mathbf{z}^t, \mathbf{z}^{v+}, \mathcal{N}^v) + \ell(\mathbf{z}^v, \mathbf{z}^{t+}, \mathcal{N}^t)\big]. \tag{13}$$

Because shots within the same drama share recurring scenes and characters, random negatives quickly become trivial and yield only coarse alignment. We therefore mine hard visual negatives from within each anchor's own drama and define the hard-negative loss as

$$\mathcal{L}_{\text{hard}} = \ell(\mathbf{z}^t, \mathbf{z}^{v+}, \mathcal{N}^{v,h}), \;\; \mathcal{N}_i^{v,h} = \text{TopK}_{K_h}\{\mathbf{z}_j^v : j \in \mathcal{C}_i^{\text{dr}}, \; s_{ij} < \eta_i\}, \tag{14}$$

where $\mathcal{C}_i^{\text{dr}}$ is the same-drama candidate set, s_{ij} is cosine similarity, and η_i excludes the top 10% most similar shots against false negatives. Combined with the bidirectional term into the final objective:

$$\mathcal{L}_{\text{rm}} = \lambda_1 \mathcal{L}_{\text{nce}} + \lambda_2 \mathcal{L}_{\text{hard}}. \tag{15}$$

Once trained, the alignment model is frozen and supplies the geometric similarity signal for both retrieval and reward throughout the framework.

4 Evaluation

We evaluate DramaDirector through seven questions: **RQ1** how it compares with end-to-end baselines; **RQ2** which components contribute to its performance; **RQ3** how discriminative text–visual alignment is for retrieval; **RQ4** how retrieval affects first-frame generation; **RQ5** how geometry retrieval behaves under modality, noise, and cross-drama controls; **RQ6** how sensitive it is to training-step choices; and **RQ7** what the end-to-end pipeline looks like qualitatively.

4.1 Experimental Settings

4.1.1 Datasets and Evaluation Protocols

DramaBoard is constructed from 35 live-action short dramas segmented into shot-level timelines with aligned keyframes and dialogue (Table 1). Auxiliary tasks, text-visual alignment, and generation follow the continuation split at 8:1:1, with 300 test videos and 600 images (Lai et al., 2026a). Judge-scored criteria use a 3-point scale $\{0, 0.5, 1\}$, while Similarity and DIS are automatic continuous metrics. Evaluation covers three dimensions: **(i) Storyboard narrative quality** (Chhun et al., 2022) – *Relevance* (alignment with context), *Coherence* (logical transitions and pacing), and *Engagement* (emotional resonance and rhythm); **(ii) Instruction following** (Feng et al., 2026) – *Faithfulness* (conformance to directives), *Similarity* (cosine similarity between Qwen3-VL embeddings of generated videos and dynamic narrative prompts), and *Purity* (absence of hallucinations); **(iii) Video intrinsic quality** (Guo et al., 2026) – *Aesthetics* (lighting, color grading, and composition), *Consistency* (stable appearance and geometry), and *DIS* (masked SSIM distance after warping adjacent 1-FPS frames with DIS optical flow; lower is better). For more details, please refer to Appendix A.5 and Appendix A.7.

Table 1: Statistics of the experimental datasets.

Task	Drama	Episodes	Shots	Dur	Samples
Continuation	35	2,807	81,283	50.2h	13,262
Shot prediction	35	2,812	27,437	16.9h	27,437
Field completion	35	2,455	6,660	4.1h	6,660
Shot ordering	35	2,437	30,534	18.7h	6,893
Reconstruction	35	2,767	58,485	35.9h	15,121

Table 2: Overall performance comparison across eight baseline settings. Best and second-best results are highlighted in bold and underlined, excluding the two variants.

Model	Storyboard Narrative Quality			Instruction Following			Video Intrinsic Quality		
	Relevance	Coherence	Engagement	Faithfulness	Purity	Similarity	Aesthetics	Consistency	DIS ↓
SkyScript	0.4850	0.4900	0.4850	0.6667	0.8333	0.5855	0.6467	0.7000	0.1753
ShoulderShot	0.5800	0.6233	0.5933	0.6933	0.9333	0.5275	0.6917	0.6933	0.1618
SkyReels-A2	0.5350	0.5333	0.5150	0.7517	0.9233	0.5626	0.5900	0.6600	<u>0.1596</u>
Dreamrunner	0.5850	0.5850	0.5683	**0.7950**	<u>0.9667</u>	0.6098	0.6067	**0.9283**	0.1942
GenMac	0.5300	0.5300	0.5067	0.7150	0.8800	0.6267	0.6033	0.7483	0.2026
MovieAgent	0.6000	0.6000	0.5550	0.7100	0.8733	0.6340	0.6283	0.7367	0.1903
UniVA	<u>0.6633</u>	<u>0.6800</u>	**0.6200**	0.6200	0.8900	0.6028	<u>0.7167</u>	0.7733	0.1799
VideoAuteur	0.5900	0.5950	0.5783	0.7450	0.9533	<u>0.6629</u>	0.6350	0.6717	0.1657
DramaDirector	**0.7230**	**0.7732**	<u>0.6032</u>	<u>0.7712</u>	**0.9752**	**0.6826**	**0.8189**	<u>0.9003</u>	**0.1549**
+ Wan 2.6	0.7817	0.7650	0.6633	0.7217	0.8850	0.6643	0.7183	0.8400	0.1881
+ NanoBanana 2	0.6800	0.6933	0.6167	0.7700	0.8967	0.6526	0.5750	0.7617	0.1663
σ	$\pm$0.0253	$\pm$0.0202	$\pm$0.0141	$\pm$0.0123	$\pm$0.0086	-	$\pm$0.0166	$\pm$0.0130	-

Table 3: Performance of ablated DramaDirector across generation and retrieval.

	Metric	Faith.	Purity	Consis.	DIS ↓
Generation	Text-only	0.6083	0.9250	0.5750	<u>0.1847</u>
	Depth-only	<u>0.6583</u>	0.8667	<u>0.6333</u>	0.2065
	Pose-only	0.6333	<u>0.9500</u>	0.5750	0.2143
	Origin	**0.7712**	**0.9752**	**0.9003**	**0.1549**

	Metric	R@10	N@10	R@20	N@20
Retrieval	w/o $\mathcal{L}_{\text{nce}}$	0.0752	0.0368	0.1198	0.0480
	w/o $\mathcal{L}_{\text{hard}}$	<u>0.0836</u>	<u>0.0416</u>	<u>0.1323</u>	<u>0.0538</u>
	w/o pose	0.0184	0.0089	0.0343	0.0129
	w/o depth	0.0013	0.0004	0.0031	0.0009
	Origin	**0.0954**	**0.0483**	**0.1429**	**0.0603**

4.1.2 Baseline Methods

We compare DramaDirector with baselines and adaptations from four categories: **(i) short-drama generation**: SkyScript (Tang et al., 2024), ShoulderShot (Zhang et al., 2025b), SkyReels-A2 (Fei et al., 2025), and Dreamrunner (Wang et al., 2026); **(ii) image generation**: ComfyAgent (Xue et al., 2025), ComfyMind (Guo et al., 2026), and Gen-Search (Feng et al., 2026); **(iii) generative agents**: GENMAC (Huang et al., 2026), MovieAgent (Wu et al., 2025), UniVA (Liang et al., 2025), and VideoAuteur (Xiao et al., 2025a); and **(iv) retrieval-augmented frameworks**: LightRAG (Guo et al., 2024), RAG-Anything (Guo et al., 2025), and VidoRAG (Zhang et al., 2026).

Implementation details of DramaDirector and the baseline methods are provided in Appendix A.11.

4.2 Overall Performance Comparison (RQ1)

In the reported comparison shown in Table 2, excluding generator-swap variants, DramaDirector scores are highest on Relevance and Coherence for storyboard narrative quality. For **Instruction Following**, the reported Similarity and Purity scores are highest for DramaDirector, while Dreamrunner remains strongest on Faithfulness. For **Video Intrinsic Quality**, the reported DramaDirector scores are highest on Aesthetics and lowest on DIS, and rank second on Consistency; Dreamrunner is strongest on Consistency. These reported values together indicate that the proposed planning and retrieval interface is competitive across narrative, instruction-following, and visual-quality criteria alike, while the generator-swap variants show that its performance depends in part on the control fidelity of the downstream generation backend. Where another baseline leads an individual criterion, the same comparison still leaves this interface competitive on the remaining reported measures.

4.3 Ablation Study (RQ2)

Figure 3 and Table 3 report the component ablations. For **Storyboard Supervision**, removing auxiliary tasks lowers coherence and temporal consistency, while the no-retrieval-reward variant shows the largest coherence drop. For **Retrieval Grounding**, removing depth reduces retrieval to near zero and removing pose also causes a substantial drop, whereas removing either contrastive-loss term has a smaller effect. For **Generation Balance**, the full depth–pose design is strongest among the reported criteria, with depth providing spatial-layout cues and pose providing body-configuration cues. These variants indicate that auxiliary supervision, the retrieval reward, and joint depth–pose grounding play complementary roles across the reported generation and retrieval criteria. In both the generation block and the retrieval block, the full setting remains ahead of every reported removal. An additional retrieval-reward factorial ablation is reported in Appendix A.12.

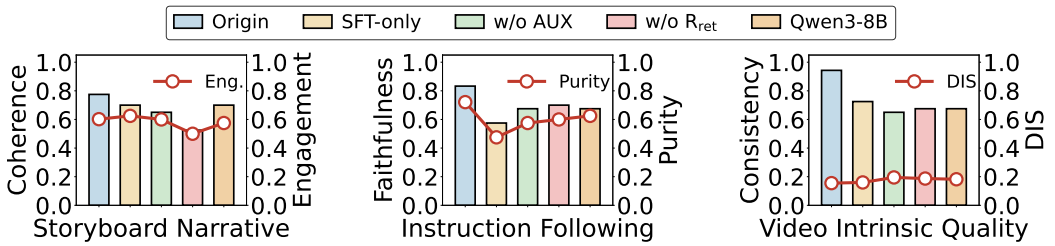

Figure 3: Tuning Performance of different ablations.

Table 4: Retrieval performance comparison.

Method	R@10	N@10	R@20	N@20
tongyi-vision-flash	0.0687	0.0395	0.1008	0.0475
tongyi-vision-plus	0.0795	0.0455	0.1130	0.0540
LightRAG	0.0781	0.0451	0.1056	0.0521
RAGAnything	0.0833	0.0472	0.1176	0.0558
VidoRAG	0.0909	0.0469	0.1241	0.0597
DramaDirector	**0.0954**	**0.0483**	**0.1429**	**0.0603**

Table 5: Hyperparameter Study on SFT steps

Step	Narrative Quality		Intrinsic Quality	
	Relevance	Engagement	Consis.	DIS ↓
40	0.6050	0.4767	0.7200	0.1938
120	0.6683	0.5284	0.7967	0.1740
240	0.6367	0.5667	0.8867	0.1643
360	0.7050	0.5867	0.8683	0.1581
426 (Ours)	**0.7230**	**0.6032**	**0.9003**	**0.1549**

4.4 Retrieval Discriminability (RQ3)

Table 4 compares DramaDirector with general-purpose vision-language encoders and retrieval-augmented methods; the structural stress tests are reported separately and do not reproduce this learned-embedding benchmark. DramaDirector has the highest score on every listed metric. The pattern matches the role of bidirectional contrastive alignment and same-drama hard-negative mining in separating compatible geometry from generic semantic similarity. VidoRAG, the closest retrieval-trained comparator, is lower on all four metrics. The larger gap at the recall cutoff, together with narrower NDCG differences, indicates that candidate coverage improves more than top-position precision, which motivates further analysis of retrieval concentration and gallery coverage. This margin over the strongest retrieval baseline holds here at both recall and NDCG cutoffs.

4.5 Quality of First-Frame Generation (RQ4)

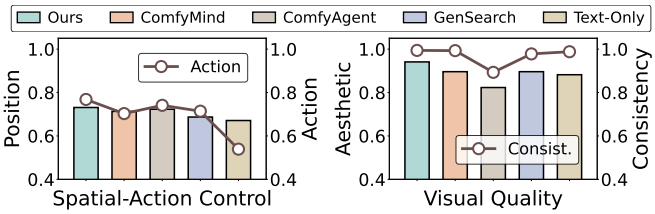

Figure 4: Quality of First-Frame Generation.

The reported Figure 4 compares retrieval-grounded first-frame generation against ComfyMind, ComfyAgent, GenSearch, and a Text-Only baseline. **Visual Quality.** The reported DramaDirector scores are highest on Aesthetics and Consistency among the compared settings, while the text-only baseline is lower on both. **Spatial–Action Control.** The reported DramaDirector scores are also highest on Position and Action, with the largest difference on Action. These values are consistent with the role of retrieved visual conditions in providing cues that are difficult to specify through text alone. **Complementarity.** Together with Table 3, the results support a division of labor in which depth primarily provides spatial-layout cues and pose primarily provides body-configuration cues.

4.6 Geometry Stress Tests and Controls (RQ5)

To separate structure from appearance and episode leakage, we score 10,000 paired shots against a candidate set of 101,334 depth–pose records, after dropping the query episode (Table 6; Appendix A.15). Matched controls beat shuffled controls on top-1 similarity and metadata agreement; noise at $\sigma = 0.25$ degrades most, and depth-only cosine is higher because its scale is more concentrated. On the gallery Recall@10 used in Table 3 and Table 4, the full depth–

Table 6: Structural retrieval stress test.

Control	Cos.	Unique	Scale	Angle
Matched	0.7130	91.26%	31.44%	72.39%
Depth	0.9060	92.02%	31.48%	71.99%
Pose	0.7010	87.26%	27.79%	71.36%
Glob. shuf.	0.6822	89.02%	24.31%	68.76%
Drama shuf.	0.6888	90.39%	25.99%	69.12%
Both, σ=0.25	0.3431	95.02%	27.56%	69.66%

pose model scores 0.0954; removing depth lowers it to 0.0013 and removing pose lowers it to 0.0184. Leave-one-drama-out over 35 dramas raises shot-scale and people-count agreement while keeping camera-motion agreement (Appendix A.14). Relative to the descriptor space, the learned depth–pose alignment therefore improves structural agreement without giving up camera-motion agreement, which is the advantage of the proposed retrieval interface in the reported cross-drama comparison.

4.7 Hyperparameter Study (RQ6)

Table 5 and Figure 5 report the training-step choices for SFT and GRPO. **SFT steps.** The reported intermediate SFT checkpoints show non-monotonic behavior: narrative scores fluctuate, consistency generally improves, and DIS decreases as training proceeds. The selected checkpoint at step 426 (Ours) has the best values among the reported checkpoints on all four metrics. **GRPO steps.** The reported curves show the best balance near the selected checkpoint at step 441 (Ours), with higher consistency and lower DIS while retaining competitive relevance and engagement. Both selected checkpoints are the reported steps that jointly raise consistency and lower DIS.

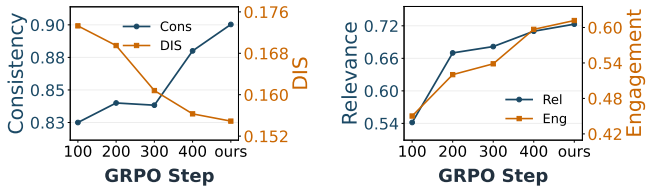

Figure 5: Hyperparameter study on GRPO steps.

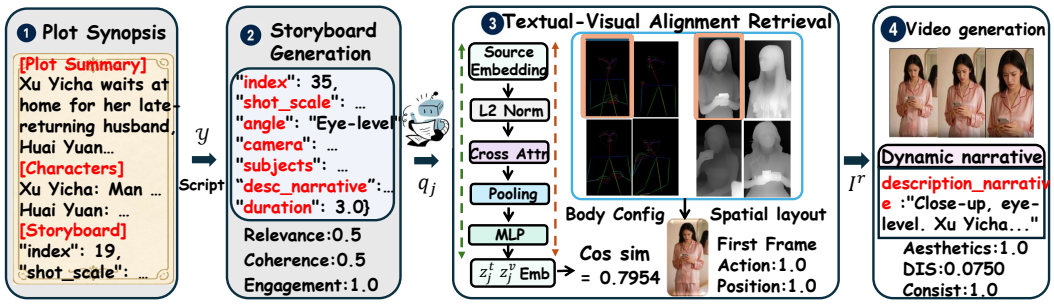

Figure 6: The end-to-end short drama generation process

4.8 Case Study (RQ7)

Figure 6 illustrates the end-to-end generation process. Given a plot summary, character profiles, and contextual storyboard history, DramaDirector first produces a structured storyboard. It then specifies cinematic fields such as "Close-up," "Eye-level," and "Static," retrieves depth–pose controls compatible with the static condition, synthesizes the first frame, and animates each shot from its dynamic narrative condition. The example shows how retrieval-grounded conditioning can connect storyboard actions to frame-level layout and subsequent video synthesis. The static condition is what the retrieved depth–pose reference grounds, while the dynamic condition is what drives the shot animation. Read in this order, the example makes the planner-to-generator interface visible from storyboard fields through the first frame to the finished shot.

5 Conclusion

This paper frames plot-to-short-drama generation as a storyboard planning problem with a missing visual interface between the planner and the generator. DramaDirector addresses this gap by retrieving appearance-reduced depth–pose references drawn from real short-drama shots as frame-level structural priors and by using visual retrievability as a policy-learning signal for the planner, with static and dynamic conditions connecting retrieval-grounded first frames to image-to-video synthesis. We further construct DramaBoard, a shot-level benchmark with aligned keyframes, dialogue, structured storyboards, and protocols spanning narrative quality, instruction following, retrieval, and video quality. The current descriptor-level diagnostics identify gallery imbalance, structural hard cases, and sensitivity to depth–pose corruption; future work should validate these effects with learned retrieval and downstream generation, audit automatic annotations, test cross-domain transfer, evaluate audiovisual synchronization, and record complete end-to-end deployment cost.

Ethical Considerations

The DramaBoard benchmark is constructed from 35 live-action short dramas. The source materials were obtained with the necessary copyrights, usage rights, and distribution permissions, and the released dataset is intended solely for academic research rather than commercial use. An anonymous human evaluation rates anonymized, randomly shuffled outputs.

AI use statement

In preparing this submission, generative AI tools were used to assist with paper editing, readability improvements, and LaTeX formatting. We reviewed and verified all AI-assisted changes and take responsibility for the final content, claims, and artifacts in this submission.

References

Alibaba Cloud. Tongyi-embedding-vision: Multimodal embedding api. https://help.aliyun.com/zh/model-studio/embedding, 2024. Accessed: 2026-05-20.

Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, et al. Onestory: Coherent multi-shot video generation with adaptive memory. *arXiv preprint arXiv:2512.07802*, 2025.

Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. Vita: An efficient video-to-text algorithm using vlm for rag-based video analysis system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2266–2274, 2024.

Gengchen Cao, Tianke He, Yixuan Liu, and RAY LC. Audience in the loop: Viewer feedback-driven content creation in micro-drama production on social media. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–25, 2026.

Cyril Chhun, Pierre Colombo, Fabian Suchanek, and Chloé Clavel. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 5794–5836, 2022.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion trans formers. *arXiv preprint arXiv:2504.02436*, 2025.

Kaituo Feng, Manyuan Zhang, Shuang Chen, Yunlong Lin, Kaixuan Fan, Yilei Jiang, Hongyu Li, Dian Zheng, Chenyang Wang, and Xiangyu Yue. Gen-searcher: Reinforcing agentic search for image generation. *arXiv preprint arXiv:2603.28767*, 2026.

Honghao Fu, Miao Xu, Yiwei Wang, Dailing Zhang, Jun Liu, and Yujun Cai. Videostir: Understanding long videos via spatio-temporally structured and intent-aware rag. *arXiv preprint arXiv:2604.05418*, 2026.

Litao Guo, Xinli Xu, Luozhou Wang, Jiantao Lin, Jinsong Zhou, Zixin Zhang, Bolan Su, and Yingcong Chen. Comfymind: Toward general-purpose generation via tree-based planning and reactive feedback. *Advances in Neural Information Processing Systems*, 38:45128–45164, 2026.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2(3), 2024.

Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. Rag-anything: All-in-one rag framework. *arXiv preprint arXiv:2510.12323*, 2025.

Chan-Wei Hu, Yueqi Wang, Shuo Xing, Chia-Ju Chen, and Zhengzhong Tu. mrag: Elucidating the design space of multi-modal retrieval-augmented generation. *arXiv preprint arXiv:2505.24073*, 2025. URL https://arxiv.org/abs/2505.24073.

Kaiyi Huang, Yukun Huang, Xintao Wang, Zinan Lin, Xuefei Ning, Pengfei Wan, Di Zhang, Yu Wang, and Xihui Liu. Filmaster: Bridging cinematic principles and generative ai for automated film generation. *arXiv preprint arXiv:2506.18899*, 2025a.

Kaiyi Huang, Yukun Huang, Xuefei Ning, Zinan Lin, Yu Wang, and Xihui Liu. Genmac: compositional text-to-video generation with multi-agent collaboration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5049–5057, 2026.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025b. URL https://arxiv.org/abs/2503.06749.

Jinxiang Lai, Zexin Lu, Jiajun He, Rongwei Quan, Wenzhe Zhao, Qinyu Yang, Qi Chen, Qin Lin, Chuyue Li, Tao Gao, et al. Visioncreator: A native visual-generation agentic model with understanding, thinking, planning and creation. *arXiv preprint arXiv:2603.02681*, 2026a. URL https://arxiv.org/abs/2603.02681.

Jinxiang Lai, Wenzhe Zhao, Zexin Lu, Hualei Zhang, Qinyu Yang, Rongwei Quan, Zhimin Li, Shuai Shao, Song Guo, and Qinglin Lu. Visioncreator-r1: A reflection-enhanced native visual-generation agentic model. *arXiv preprint arXiv:2603.08812*, 2026b. URL https://arxiv.org/abs/2603.08812.

Hang Lei, Shengyi Zong, Zhaoyan Li, Ziren Zhou, and Hao Liu. Beyond direct generation: A decomposed approach to well-crafted screenwriting with llms. *arXiv preprint arXiv:2510.23163*, 2025. URL https://arxiv.org/abs/2510.23163.

Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Liu Si. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*, 2025a.

Jiachen Li, Qian Long, Jian Skyler Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Wang. T2v-turbo-v2: Enhancing video model post-training through data, reward, and conditional guidance design. In *International Conference on Learning Representations*, volume 2025, pp. 92279–92305, 2025b.

Zhengyang Liang, Daoan Zhang, Huichi Zhou, Rui Huang, Bobo Li, Yuechen Zhang, Shengqiong Wu, Xiaohan Wang, Jiebo Luo, Lizi Liao, et al. Univa: Universal video agent towards open-source next-generation video generalist. *arXiv preprint arXiv:2511.08521*, 2025.

Mingyang Mao, Mariela M. Perez-Cabarcas, Utteja Kallakuri, Nicholas R. Waytowich, Xiaomin Lin, and Tinoosh Mohsenin. Multi-rag: A multimodal retrieval-augmented generation system for adaptive video understanding. *arXiv preprint arXiv:2505.23990*, 2025.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Vader: Video diffusion alignment via reward gradients. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.08737.

Jing Tang, Quanlu Jia, Yuqiang Xie, Zeyu Gong, Xiang Wen, Jiayi Zhang, Yalong Guo, Guibin Chen, and Jiangping Yang. Skyscript-100m: 1,000,000,000 pairs of scripts and shooting scripts for short drama. *arXiv preprint arXiv:2408.09333*, 2024. URL https://arxiv.org/abs/2408.09333.

Maria Teleki, Vedangi Bengali, Xiangjue Dong, Sai Tejas Janjur, Haoran Liu, Tian Liu, Cong Wang, Ting Liu, Yin Zhang, Frank Shipman, et al. A survey on llms for story generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13954–13966, 2025.

Zun Wang, Jialu Li, Han Lin, Jaehong Yoon, and Mohit Bansal. Dreamrunner: Fine-grained compositional story-to-video generation with retrieval-augmented motion adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10503–10511, 2026.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*, 2025.

Junfei Xiao, Feng Cheng, Lu Qi, Liangke Gui, Yang Zhao, Shanchuan Lin, Jiepeng Cen, Zhibei Ma, Alan Yuille, and Lu Jiang. Videoauteur: Towards long narrative video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19163–19173, 2025a.

Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain cinema: Towards short movie generation. *arXiv preprint arXiv:2507.18634*, 2025b. URL https://arxiv.org/abs/2507.18634.

Xiangyuan Xue, Zeyu Lu, Di Huang, Zidong Wang, Wanli Ouyang, and Lei Bai. Comfybench: Benchmarking llm-based agents in comfyui for autonomously designing collaborative ai systems. In *Proceedings of the computer vision and pattern recognition conference*, pp. 24614–24624, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4210–4220, 2023.

Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. Stage: Storyboard-anchored generation for cinematic multi-shot narrative. *arXiv preprint arXiv:2512.12372*, 2025a.

Wenchuan Zhang, Jingru Guo, Hengzhe Zhang, Penghao Zhang, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. Patho-agenticrag: towards multimodal agentic retrieval-augmented generation for pathology vlms via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29921–29929, 2026.

Yuang Zhang, Junqi Cheng, Haoyu Zhao, Jiaxi Gu, Fangyuan Zou, Zenghui Lu, and Peng Shu. Shouldershot: Generating over-the-shoulder dialogue videos. *arXiv preprint arXiv:2508.07597*, 2025b.

A APPENDIX

A.1 BASELINE METHODS

This appendix outlines the backbones, baselines, and format-based adaptations used for end-to-end short-drama generation. We employ Vidu Q3 Turbo and Seedream v5 Lite as our primary video and image backbones, and include Wan 2.6 and NanoBanana 2 to assess behavior under generator swaps. We compare with four categories of methods: (1) short-drama formatting and generation, (2) image generation, (3) generative agents, and (4) retrieval augmentation.

Short-Drama Generation Methods

- **SkyScript** (Tang et al., 2024): We instantiate this comparison by following its script–shooting-script format: the input plot is serialized into a shooting-script-style shot plan, and the resulting shot descriptions are passed to the same downstream image and video generation interface used by the other generation baselines.
- **ShoulderShot** (Zhang et al., 2025b): Generates over-the-shoulder dialogue videos using dual-shot generation and looping mechanisms for spatial and character consistency.
- **SkyReels-A2** (Fei et al., 2025): An elements-to-video (E2V) framework assembling visual elements into scenes via image-text joint embedding for global coherence.
- **Dreamrunner** (Wang et al., 2026): Combines LLM layout planning, retrieval-augmented motion adaptation, and spatial-temporal control for story-to-video synthesis.

Image Generation Frameworks

- **ComfyMind** (Guo et al., 2026): A collaborative AI system for general-purpose generation, featuring a Semantic Workflow Interface and search tree planning with localized feedback.
- **Gen-Search** (Feng et al., 2026): A search-augmented agent that utilizes multi-hop reasoning to search online for textual knowledge and visual references for grounded image generation.
- **ComfyAgent** (Xue et al., 2025): An LLM-based framework that autonomously designs collaborative multi-agent AI systems by generating executable workflows.

Generative Agentic Frameworks

- **MovieAgent** (Wu et al., 2025): Employs hierarchical CoT to organize scripts, scenes, and cinematography for long-form video generation.
- **UniVA** (Liang et al., 2025): Decomposes user intentions into modular steps for versatile video understanding, editing, and generation.
- **VideoAuteur** (Xiao et al., 2025a): Synthesizes long narrative videos by planning captions, actions, and visual states for coherent generation.
- **GENMAC** (Huang et al., 2026): Iteratively refines complex text-to-video prompts through multi-agent design, generation, and redesign.

Retrieval-Augmented Frameworks

- **Tongyi-Embedding-Vision** (Alibaba Cloud, 2024): A multimodal embedding API from Alibaba Cloud that projects images and text into a unified vector space for efficient cross-modal retrieval and semantic alignment.
- **LightRAG** (Guo et al., 2024): Combines structural indexing and dual-level graph retrieval for efficient knowledge extraction.
- **RAG-Anything** (Guo et al., 2025): Enables versatile multimodal retrieval (e.g., text, images, tables) via cross-modal graph reasoning.
- **VidoRAG** (Zhang et al., 2026): A multimodal RAG framework featuring joint text-image search and multi-turn reasoning.

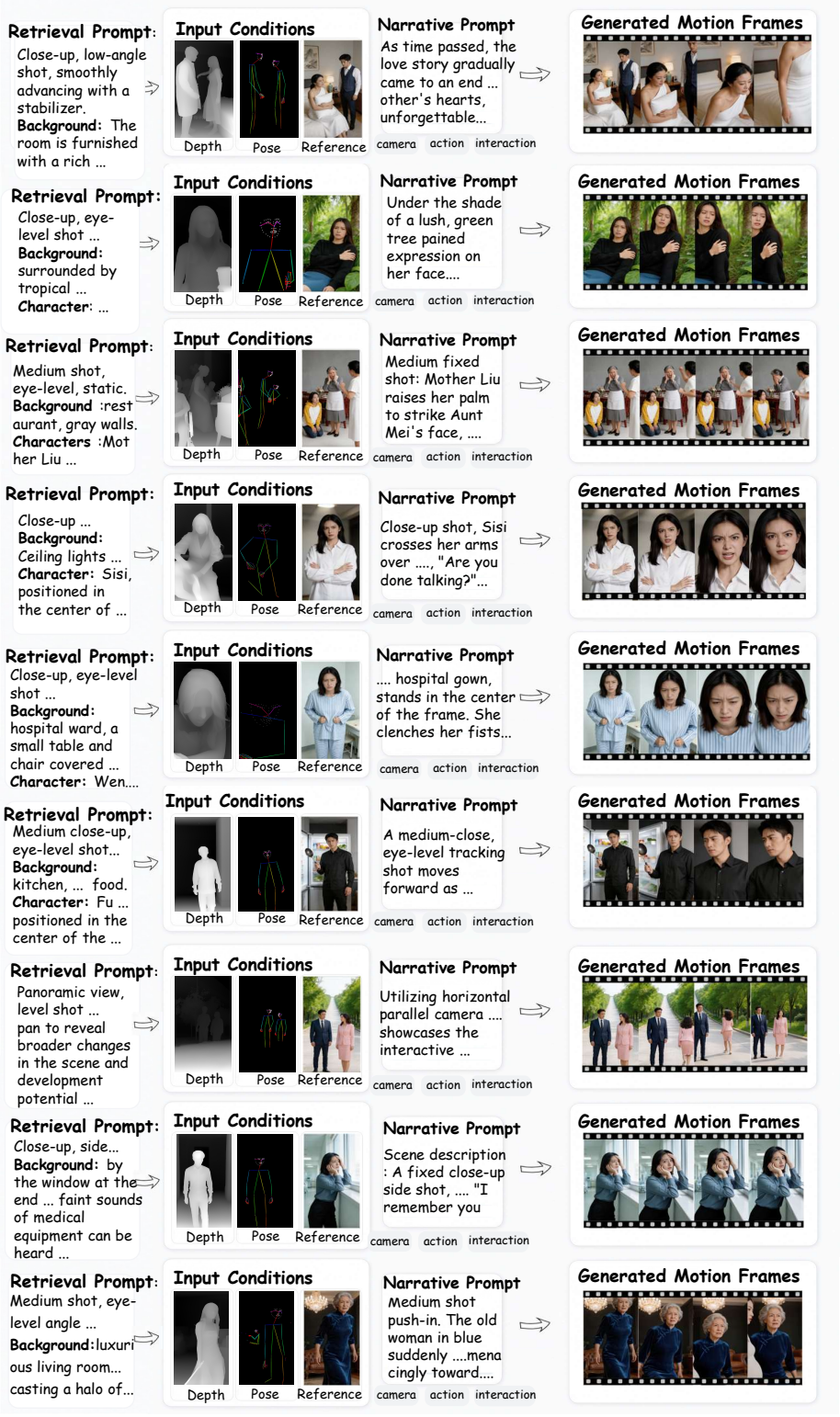

Figure 7: Comparison of DramaDirector, image-to-video, and text-to-video workflows

A.2 Generation Workflow Comparison

To illustrate the intended role of structural retrieval, we compare three workflows between storyboard text and video generation in Figure 7: DramaDirector, image-to-video, and text-to-video. DramaDirector retrieves depth and pose conditions from a short-drama storyboard library, providing explicit cues for shot composition, spatial layout, and coarse character motion that are not available in text-only generation. Relative to RGB references, depth and pose reduce appearance variation and expose structural cues more directly. The examples are qualitative illustrations of this interface; systematic comparisons are reported in the main evaluation.

A.3 Planner Generalization Across Dramas and Domains

We evaluate cold-start storyboard generation from 60 English-language movie summaries in the CMU Movie Summary Corpus. We parse the generated storyboards into shots, sample 200 shot-level evaluations, and pass each shot to a text judge together with its source summary and adjacent shots. The judge scores Relevance, Coherence, and Engagement on a $\{0, 0.5, 1\}$ three-point scale. Across these 200 shots, the planner obtains mean scores of 0.6250, 0.7125, and 0.5625. Coherence is the strongest cross-domain dimension, while Engagement remains comparatively lower, consistent with the difficulty of transferring emotional pacing and short-drama rhythm to unseen English-language movie summaries.

A.4 Hyperparameter Study

We examine two hyperparameter groups that govern retrieval alignment. **Depth–pose fusion weights** (Table 7) show that extreme weighting degrades retrieval. While a slightly pose-emphasized setting (0.4/0.6) achieves the best overall retrieval metrics, its performance gain over equal fusion (0.5/0.5) is marginal (e.g., only a 0.24 difference in N@10). During downstream video synthesis, decreasing the depth weight below 0.5 was associated with more frequent deviations in broad scene layout and camera composition. We therefore adopt equal weights (0.5/0.5) as a balance between multi-person pose configuration and spatial structure. **Contrastive loss weights** (Figure 8) show that removing or over-weighting either the standard or hard-negative term degrades retrieval under the evaluated settings, suggesting that inter-sample negatives and same-drama hard negatives provide complementary supervision.

Table 7: Study on depth and pose weights.

D/P	R@10	N@10	R@20	N@20
0.1/0.9	7.12	3.45	11.58	4.57
0.2/0.8	8.43	4.20	14.04	5.62
0.3/0.7	4.51	2.11	7.85	2.95
0.4/0.6	**9.74**	**5.07**	**14.48**	**6.26**
0.5/0.5*	9.54	4.83	14.29	6.03

A.5 Benchmarking

The collection covers diverse themes, including time-travel, costume drama, fantasy, urban relationships, family conflicts, suspense, and revenge. The dramas differ substantially in scene types, character configurations and spatial relations, shot scales and camera motions, dialogue density, and narrative pacing; the benchmark therefore contains variation in both story themes and cinematographic organization, enabling evaluation of planning and generation under heterogeneous short-drama conditions.

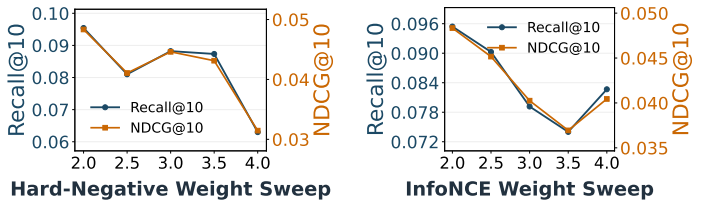

Figure 8: Balancing Negative and Hard Negative Losses.

We construct the benchmark from raw short-drama videos through a staged pipeline, producing a shot-level storyboard corpus for both supervised training and downstream generation. The released shot inventory contains 2,826 episodes and 101,763 shot records; the Continuation subset used for the reported task splits contains 2,807 episodes and 81,283 shots (Table 1). Unless otherwise specified, the multimodal analysis, transcript correction, and storyboard annotation stages use Qwen3.5-Plus as the annotation backbone.

Shot preprocessing Each episode is segmented into shots with one representative keyframe. A boundary at time t is kept if

$$\delta_t > \theta_{\text{cut}}, \qquad \Delta_t \geq L_{\min}, \tag{16}$$

where δ_t denotes the inter-frame visual difference score, Δ_t is the resulting segment length, $\theta_{\text{cut}} = 30.0$ is the cut threshold, and $L_{\min} = 15$ frames is the minimum shot length. We retain shot indices, timestamps, durations, keyframe paths, and source-video paths, which are reused in later alignment and annotation stages.

Depth–pose preprocessing For each representative keyframe, we extract the geometric controls used by retrieval and first-frame conditioning. Depth maps are produced with Depth Anything V2 (Yang et al., 2024), using the ViT-L encoder checkpoint and an input size of 518, then min–max normalized to an 8-bit control image. Pose maps are produced with DWPose (Yang et al., 2023), using the ONNX whole-body pipeline with YOLOX-L human detection and the `dw-ll_ucoco_384` pose estimator. The rendered pose control includes body, hand, and face keypoints on a blank canvas, so the downstream gallery stores geometry rather than appearance.

Gallery construction and split policy The depth–pose gallery is constructed from the available short-drama episodes and is used as a contextual visual prior rather than as a target storyboard label. To prevent episode-level leakage during evaluation, we remove all shots from the source episode of each test sample before retrieval. Thus, generated storyboards cannot retrieve depth–pose references from the same episode as the held-out target, and the held-out storyboard annotations are used only for evaluation. Retrieval instead operates over depth and pose references from other episodes, supporting reuse of cinematographic structure while reducing the risk of memorizing episode-specific layouts. This setting retains a practical short-drama geometry gallery while enforcing separation between test targets and retrieval evidence.

Checkpoint provenance for learned retrieval The externally located reward-model checkpoint records a 768-dimensional Tongyi input space, a 2048-dimensional projection space, and test-set metadata of Recall@1/5/10/20 equal to 0.0153/0.0576/0.0954/0.1429, with mean positive score 0.6127. The recorded Recall@10 and Recall@20 match the $DramaDirector$ row in Table 4.

Episode-level semantic annotation We next run a full-video multimodal analysis pass to extract a character list, a plot summary, and a global dialogue reference. These episode-level conditions provide cross-shot semantic grounding for later annotation and off-screen speaker attribution, while also improving consistency in character naming and narrative interpretation across the episode.

Shot-level transcript correction We align speech to shot boundaries, run ASR on each shot, and apply an LLM-based correction step conditioned on the character list, plot summary, and dialogue reference. This step fixes recognition errors while preserving shot indices and empty transcripts. It is particularly useful for short-drama speech with colloquial expressions, aliases, omitted subjects, and off-screen utterances.

Shot-level storyboard annotation Given the keyframe, shot duration, corrected transcript, and episode-level context, we prompt a multimodal model to produce one structured storyboard object per shot, as shown in 3. Each object includes camera fields, subjects, background, narrative, dialogue, speaker, emotion, and duration. The prompt enforces one-to-one frame-to-shot correspondence, avoids unverifiable psychological descriptions, and resolves off-screen speech to the speaker. This stage converts multimodal signals into a consistent representation suitable for retrieval and generation.

Automatic schema audit As a mechanical check on the fixed data release, we audit all 101,763 shot records. Four records contain a misspelled variant of `description_narrative`, one record lacks an `index`, and two episode files have a non-contiguous index sequence. Among 129,382 subject entries, 708 omit `expression` and 5 omit `clothing`; 2,943 shot records have an empty subject list. These counts describe schema completeness only and do not estimate semantic annotation accuracy; the stratified manual audit below provides semantic verification at the reported sample scale.

Table 8: Manual annotation reliability based on 500 sampled storyboard shots.

Audit item	Mean [95% CI]	ICC(2,1)	ICC(2,3)
Overall core fields	3.72 [3.66, 3.78]	0.74	0.895
Frame–shot correspondence	4.56 [4.52, 4.60]	0.90	0.964
Shot boundary and duration	4.16 [4.11, 4.21]	0.85	0.944
Subject identity and count	4.14 [4.09, 4.19]	0.83	0.936
Scene and background	3.80 [3.74, 3.86]	0.63	0.836
Camera fields	3.77 [3.71, 3.83]	0.60	0.818
Action and narrative	3.87 [3.81, 3.93]	0.70	0.875
Dialogue and speaker	3.93 [3.87, 3.99]	0.78	0.914

Multi-task SFT We organize storyboard sequences into a multi-task SFT centered on continuation, with one example shown in Figure 4. Auxiliary tasks include single-shot prediction, field completion, shot ordering, and summary-guided reconstruction, providing supervision for local continuity, structured completion, temporal coherence, and cross-shot semantic consistency. The continuation subset is stored as `system`/`user`/`assistant` triplets, and all tasks are serialized with a chat template for training.

TTS usage In the application pipeline, the `dialogue`, `speaker`, and `emotion` fields are consumed by a TTS module to synthesize speaker-aware expressive speech. Thus, the structured storyboard serves both as a supervision target and as an intermediate representation for end-to-end short-drama production.

Annotation quality control The benchmark annotations are produced by an automated pipeline: shots are segmented from the raw videos, transcripts are corrected by ASR followed by an LLM, and each storyboard object is generated by a multimodal (VLM) model. Because this pipeline is automatic, correctness is not guaranteed by construction. We therefore verify the annotation quality with a stratified manual evaluation. We sample 500 shots across dramas, episodes, and shot types. Fifty student volunteers with basic video-understanding training participated after a shared calibration session. Each shot was independently rated by three volunteers, and each volunteer reviewed 30 shots, yielding 1,500 shot-level ratings. The volunteers rated frame–shot correspondence, shot boundaries and duration, subject identity and count, scene and background descriptions, camera fields, action and narrative descriptions, dialogue content, and speaker attribution on a 5-point Likert scale, where 5 denotes accurate, specific, and verifiable; 4 denotes accurate with only minor imprecision; 3 denotes acceptable but underspecified; 2 denotes meaningful errors or omissions; and 1 denotes unusable or contradictory.

For each field, we report the mean with its 95% confidence interval and two-way random-effects ICCs. $ICC(2,1)$ describes single-rater reliability, while $ICC(2,3)$ describes the reliability of the average of the three raters assigned to each shot. The mean scores range from 3.72 to 4.56, and all fields have average-rater ICC above 0.80. The most subjective fields, scene/background and camera fields, have the lowest single-rater reliability, whereas structural fields such as frame–shot correspondence and shot boundaries have the highest. These results indicate that the automatically produced annotations are sufficiently reliable for training and evaluation when three raters are averaged.

A.6 Reward and GRPO Specification

This subsection specifies the GRPO reward and optimization settings used in the reported runs, including values that would otherwise be left to library defaults.

Visual-query serialization Before text embedding, each generated shot is converted to the same appearance-reduced format used to train the alignment model. This serialization is the operational definition of q_j: the query concatenates shot scale, camera angle, camera motion, background, and, for each subject, screen position, clothing, action, and expression. Character names, dialogue, speaker, emotion, and duration belong to d_j and are excluded from the retrieval query. Non-string values are recursively converted to text, a dictionary-valued subject is wrapped as a one-element list, and an invalid subject container becomes an empty list.

Reward components and abnormal handling Listing 9 gives the complete aggregation behavior. The format score before rescaling is

$$R_{\text{fmt}} = \frac{1}{N}\sum_{j=1}^{N}\left(0.6P_j + 0.3C_j + 0.1S_j\right), \tag{17}$$

where P_j is key coverage over the 11 required shot fields; C_j is non-empty coverage over shot scale, angle, motion, subjects, background, and narrative description; and S_j is average subfield coverage over subject name, gender, clothing, position, action, and expression. The implementation aggregates the signed term $\widetilde{R}_{\text{fmt}} = 2R_{\text{fmt}} - 1$. Because an empty subject list receives neither the subjects-content credit nor subject-subfield credit, an otherwise complete environment-only shot has $R_{\text{fmt}} = 0.85$; this is an implementation detail rather than a claim that such shots are semantically invalid. The format score is computed on the parsed structure before the sanitization used for visual-query construction.

Full video-readiness judge instruction The training judge is instructed to act as a professional AI short-drama director and to assess each ordered storyboard sequence as a whole. It assigns 20% to each of the following: (1) action concreteness—whether actions, poses, and behaviors are sufficiently explicit for controllable generation; (2) visual richness—whether shot scale, space, expression, clothing, and lighting are concrete, without penalizing justified shot–reverse-shot or same-scene repetition; (3) transition coherence—whether shot transitions and narrative changes are natural, allowing non-realistic settings when internally consistent; (4) emotional narration—whether emotion, dialogue rhythm, and dramatic tension are supported, without penalizing dialogue-free sequences; and (5) short-drama pacing—whether 1–5 second shot rhythm and conflict progression are appropriate. The overall score is the arithmetic mean in $[0, 1]$. The judge must return only a JSON array with exactly one object per sequence, using keys `index`, `score`, and `reasoning`; indices must be complete and non-repeating, and each reason is limited to 200 Chinese characters. Thinking output is disabled. The released default judge is Qwen3.5-Flash, with four sequences per request, three retries with exponential backoff, and an in-memory cache for repeated narrative sequences. Judge temperature and maximum output tokens follow the provider defaults.

GRPO configuration audit Table 9 records the training defaults used for the reported runs.

Table 9: GRPO configuration. SFT values are listed separately where they differ.

Setting	SFT	GRPO
Group size	–	4
Max prompt / completion	8,192 / –	4,600 / 2,048
Temperature / top-p	–	0.9 / 0.9
Repetition penalty	–	1.2
Learning rate	5×10^{-6}	5×10^{-7}
Per-device batch / accumulation	16 / 2	16 / 2
Epochs	3	1
Warmup ratio	0.01	0.05
Weight decay	10^{-3}	10^{-2}
Optimizer / schedule	AdamW / linear	fused AdamW / cosine
Precision / max grad norm	BF16 / 1.0	BF16 / 1.0
Clipping epsilon	–	0.2
Reward weights	–	$\alpha : 0.7 \to 0.5$, $\beta : 0.5$, $\gamma : 0.1$
Retrieval top-k	–	1

The GRPO clipping coefficient is set explicitly. A fractional warmup value of 0.05 is treated as a warmup ratio; non-integer values greater than or equal to one are rejected.

As a boundary check for the documented abnormal-output paths, we ran a fixed-seed property audit over 258 malformed or partially populated shot lists (including 28 cases with non-finite numeric values), five empty/non-JSON or truncated completion strings, and seven warmup inputs. All extracted format scores remained finite in $[0, 1]$; truncated-array normalization retained only complete objects, and non-integer warmup values at least one were rejected. This check does not call the embedding or judge APIs.

A.7 MULTIMODAL JUDGE PROMPT

To rigorously evaluate artifacts, we combine decomposed judge prompts (Figures 6 and 7) with automatic Similarity and DIS measurements. Judge prompts require metric-wise scores with evidence-grounded rationales, avoiding the ambiguity of holistic ratings and better capturing the multidimensional nature of short-drama generation, where quality depends on visual fidelity, prompt grounding, spatial layout, temporal execution, and narrative continuity.

For **Image Quality Assessment**, we evaluate each image along six dimensions: *aesthetics*, *consistency*, *faithfulness*, *attribute binding*, *position*, and *action*. This separation prevents visually appealing but semantically incorrect images from receiving inflated scores. Specifically, *faithfulness* measures core semantic alignment, *attribute binding* and *position* verify correct modifier assignment and spatial relations, and *action* assesses whether the frame captures a plausible decisive moment.

For **Video Quality**, we evaluate three axes. **Storyboard Narrative Quality** (*Relevance*, *Coherence*, *Engagement*) assesses text-only continuation, so narrative flaws cannot be hidden by appealing visuals. **Instruction Following** (*Faithfulness*, *Similarity*, *Purity*) verifies whether the video realizes the storyboard and dynamic narrative prompt while avoiding unprompted content. **Intrinsic Quality** (*Aesthetics*, *Consistency*, *DIS*) measures visual fidelity and temporal stability independent of textual alignment; DIS is the masked SSIM distance after DIS-flow warping, where lower values indicate smoother adjacent-frame transitions.

All judge-scored metrics use a $\{0, 0.5, 1\}$ scale, where 1 is reserved for complete and unambiguous satisfaction. Ambiguous or occluded details receive no credit, and character identity and story continuity remain explicit evaluation criteria.

A.8 BACKBONE INTERFACES

DramaDirector uses commercial generation backbones only through explicit input–output interfaces. We do not claim novelty in these backbones; the reproducible component of our framework is how structured storyboards, retrieved depth–pose references, and dynamic narrative fields are converted into conditioning inputs for them.

Embedding extraction For retrieval and reward modeling, we use a frozen Tongyi-Embedding-Vision encoder to embed three inputs: the static visual condition q_j, the retrieved depth map, and the retrieved pose map. Text inputs are submitted as text records, while depth and pose controls are submitted as base64-encoded images. The resulting vectors are saved as per-shot arrays and then consumed by our trainable projection, fusion, contrastive alignment, and GRPO reward modules. Thus, Tongyi provides only fixed base embeddings; retrieval scoring and policy optimization are performed by our own alignment model.

First-frame generation The first frame is not generated from the conditioning prompt in Figure 5 alone. For each shot, we first serialize the storyboard fields, including camera scale, camera angle, camera motion, subjects, background, and narrative description, into a shot-level prompt. Dialogue content is removed from this visual prompt because speech is handled by the dynamic narrative condition. We then append explicit conditioning instructions that bind the uploaded images to their roles: character reference images, if any, appear first; the retrieved depth map is uploaded next and provides cues for scene hierarchy, perspective, and spatial layout; the retrieved pose map is uploaded last and provides cues for coarse body configuration and limb orientation. The first frame is generated with the Seedream-v5-lite image-to-image interface at a vertical 9:16 resolution.

Image-to-video generation After first-frame synthesis, each shot is animated with the Vidu-q3-turbo image-to-video interface. The submitted payload contains the generated first frame, the dynamic narrative prompt derived from d_j, the shot duration, the target resolution, and the audio flag. In our experiments, video generation uses the first frame as the sole visual anchor and keeps audio generation disabled, so the visual comparison isolates storyboard-to-video realization rather than TTS quality.

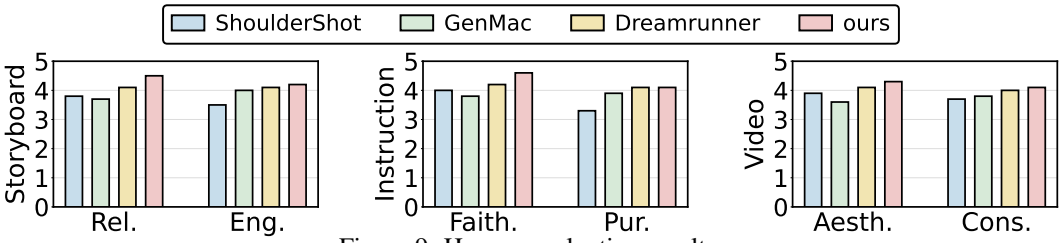

Figure 9: Human-evaluation results.

Execution records The generation interface is designed to store the exact submitted prompt, uploaded control-image order, local paths of depth and pose references, selected character references, duration, resolution, task identifier, returned URL, and downloaded output path.

A.9 Human Evaluation

We evaluate the perceptual quality of generated short-drama videos on 30 cases, each containing 4–5 storyboard shots. Anonymized outputs from DramaDirector and three baselines (ShoulderShot, GenMac, and Dreamrunner) were rated by 41 raters on a 5-point Likert scale. The dimensions are *Relevance* and *Engagement* for storyboard narrative quality, *Faithfulness* and *Purity* for instruction following, and *Aesthetics* and *Consistency* for intrinsic video quality; scores from 1 to 5 correspond to very poor, poor, fair, good, and excellent quality. Outputs were anonymized and randomly shuffled. The rubric does not isolate relative depth, position, pose, or occlusion, so the study measures overall perceived quality rather than geometry-specific correctness. Table 10 reports the mean, 95% CI, and single-rater ICC(3,1) for DramaDirector. It has the highest mean on every metric. Its ICC(3,1) lies in $[0.641, 0.749]$, and the 41-rater average ICC(3,41) lies in $[0.987, 0.992]$.

Table 10: Human-evaluation CI and ICC(3,1).

Metric	Mean [95% CI]	ICC
Relevance	4.6 [4.44, 4.76]	0.641
Engagement	4.4 [4.21, 4.59]	0.695
Faithfulness	4.6 [4.42, 4.78]	0.679
Purity	4.3 [4.11, 4.49]	0.692
Aesthetics	4.5 [4.28, 4.72]	0.749
Consistency	4.3 [4.11, 4.49]	0.674

A.10 Computational Efficiency Discussion

DramaDirector uses an 8B planner, and gallery depth–pose embeddings can be precomputed so that online retrieval requires only query encoding and nearest-neighbor search. Listing 10 reports planner runtime measurements, kept separate from the end-to-end generation results. The 80-completion, 680-completion, and 200-prompt four-seed runs are planner diagnostics, and the depth and pose rows measure low-level descriptor extraction from already generated control images, not Depth Anything or DWPose inference.

On the audited server, the Qwen3-8B model directory occupies approximately 16 GB. The released adapter and reward-model files are 174.7 MB and 153.3 MB, respectively, and the extracted-control directory occupies approximately 15 GB. These are storage measurements, not runtime memory. The per-shot API cost and end-to-end latency of the released pipeline are reported in Tables 11 and 12.

For local nearest-neighbor search over 230-dimensional fused descriptors, 100 warm repeats with a batch of 128 queries give p50/p95 batch latencies of 0.078/0.107 ms, 0.136/0.138 ms, and 0.190/0.204 ms for 10%, 50%, and 100% galleries (10,133, 50,667, and 101,334 records). These numbers measure the local nearest-neighbor search stage.

A.10.1 API Cost and End-to-End Latency

To quantify commercial API cost, we decompose the released pipeline into image generation, image-to-video generation, and LLM-based scoring or embedding. Under current public list prices, converted at 7.2 CNY per USD, the per-shot API cost is \$0.30–\$0.31. Image-to-video generation accounts for about 80% of this cost, image generation for about 19%, and LLM scoring or embedding together stay under 1%.

Table 11: Released per-shot API cost.

Item	Cost (USD)
Shot first frame	0.042
Amortized character references	0.018
Image-to-video synthesis	0.243
Judge LLM (generation quality)	<0.0005
Judge LLM (image/video evaluation)	<0.001
Text embedding	<0.0005
Total	0.30–0.31

Table 12: End-to-end latency of the released per-shot and batch pipeline.

Stage	Latency
Judge LLM (generation quality)	<1 s/shot
Seedream first frame	20–40 s/image
Vidu q3-turbo	60–180 s/shot
Judge LLM (image/video evaluation)	10–30 s/shot
Single-shot pipeline	1.5–4.2 min
Batch pipeline	8–12 min

End-to-end latency is dominated by asynchronous video synthesis. A single shot completes sequentially in 1.5–4.2 minutes, the sum of the stage ranges below, and a batch completes in 8–12 minutes under default concurrency. Tables 11 and 12 summarize the stage-level breakdown.

A.11 Implementation Details

All methods follow their original specifications where applicable. To isolate planning and grounding, all baselines and DramaDirector use the same downstream generation backbones—Seedream-v5-lite for image generation and Vidu-q3-turbo for video synthesis—except for the explicitly marked generator-swap variants. DramaDirector uses Qwen3-8B as its planner backbone. For the agentic baselines, we use Qwen3.5-Flash as the planner backbone, which is larger than the Qwen3-8B planner used by DramaDirector. This setup gives those baselines a stronger planning backbone and makes the comparison conservative with respect to planner capacity. For text–visual alignment training, we set $\lambda_1 = \lambda_2 = 2.0$, use a learning rate of 1×10^{-2}, sample 50 negatives and 54 hard negatives, and stop early at step 3. For schema-constrained supervision, the LoRA rank is 16, the learning rate is 5×10^{-6}, the optimizer is AdamW (the released SFT default is `adamw_8bit`), the warm-up ratio is 0.01, the weight decay is 1×10^{-3}, and training runs for 3 epochs. For policy optimization, the group size is 4, the maximum prompt/completion lengths are 4,600/2,048, the per-device batch size and gradient accumulation are 16/2, and the learning rate is 5×10^{-7}. The released GRPO configuration uses BF16, fused AdamW, cosine decay, warm-up ratio 0.05, weight decay 1×10^{-2}, one epoch, temperature 0.9, top-p 0.9, repetition penalty 1.2, and clipping coefficient 0.2. Advantages are normalized across the G completions in each group, as in Section 3.3. Qwen3.5-Flash provides the video-readiness reward, while Qwen3.5-Plus is used for final evaluation. We use tongyi-embedding-vision-flash-2026-03-06 for multimodal encoding. With released defaults $(\alpha_0, \alpha_{\text{end}}, \beta_0, \beta_{\text{end}}, \gamma) = (0.7, 0.5, 0.5, 0.5, 0.1)$, the retrieval weight decreases while the video weight remains constant; do not call this schedule beta annealing unless the run configuration differs. The per-shot API cost and end-to-end latency of the released pipeline are summarized in Appendix A.10.1.

A.12 Retrieval-Reward Factorial Ablation

As a factorial for the retrieval-reward component, the Qwen3-8B planner is initialized from the supervised checkpoint and trained under four reward configurations. Generated storyboards are then scored against the full episode gallery with the frozen learned reward model (Table 13). These Recall and NDCG numbers answer a different question from Table 4. That table measures how often a gold static condition retrieves its own keyframe geometry, which is the protocol behind Recall@10 of 0.0954 and Recall@20 of 0.1429. Here the query is a continuation the planner has just written, and a hit means the frozen model places the target episode's cached geometry in the top ranks of the full multi-episode gallery. The two protocols therefore sit on different absolute scales. Within this generated-query protocol, the full depth–pose retrieval reward is the strongest of the four training settings.

A.13 Gallery Composition and Diversity

The 101,763-shot gallery contains multiple camera and staging categories, but their frequencies are highly skewed. The five most frequent shot-scale labels account for 97.81% of records, the level-angle label accounts for 81.98%, and fixed-camera motion accounts for 93.43%. People counts

Table 13: Retrieval-reward factorial ablation.

Variant	R@10	N@10	R@20	N@20
SFT-only	0.0154	0.0057	0.0308	0.0098
Full (depth+pose retrieval reward)	**0.0190**	**0.0090**	**0.0332**	**0.0122**
Depth-only retrieval reward	0.0042	0.0021	0.0125	0.0042
Pose-only retrieval reward	0.0048	0.0019	0.0096	0.0030

Table 14: Dominant gallery categories over 101,763 shot records. Entropy is normalized by the observed category count.

Field	Most frequent categories	Top category	Entropy
Shot scale	Medium close-up, close-up, close, medium, wide	25.80%	0.451
Camera angle	Level, side, high, over-the-shoulder, low	81.98%	0.216
Camera motion	Static, tracking, pan, push-in	93.43%	0.086
People count	1, 2, 3, 0, 4+	72.45%	0.506

are similarly concentrated: one- and two-person shots account for 92.98%. At the drama level, a mutually exclusive dataset-oriented coding places 82.9% of titles in three modern-urban groups, with modern romance/marriage and family melodrama alone accounting for 42.9% (Table 17). Weighting by shot records rather than titles raises the modern romance/marriage/family share to 49.7%, while the single wilderness-survival title contributes only 1.4% of the gallery. The normalized entropies of shot scale, camera angle, camera motion, and people count are 0.451, 0.216, 0.086, and 0.506, respectively (Table 14). Thus, the gallery provides broad episode coverage but a narrow empirical prior over some cinematographic controls and themes; balancing and concentration diagnostics are reported in Appendix A.15.

A.14 Additional Geometry Stress Results

The learned-alignment and leave-one-drama-out results that accompany the main-text structural retrieval stress test are reported below.

A.15 Gallery Balancing and Sequence-Diversity Diagnostics

Table 15: Leave-one-drama-out agreement.

Metric	Descriptor	Learned
Shot-scale agreement	30.5%	37.0%
Camera-angle agreement	70.6%	72.5%
Camera-motion agreement	88.0%	88.6%
People-count agreement	65.4%	71.7%
Distinct dramas among top-1	33.74/34	34/34

Thematic composition We issued exact-title queries for all 35 directories; only five titles yielded pages usable for broad topic-level corroboration, while the remaining results were ambiguous, unrelated, or too sparse. The web queries therefore supplement rather than replace the released plot summaries. We assign each of the 35 drama titles one mutually exclusive primary theme using the released episode-level plot summaries. Web search is used only to cross-check terminology, representative plots, and possible same-title works. These searches confirm the broad thematic cues but also expose substantial title ambiguity. Table 17 is therefore a dataset-oriented coding rather than an official platform label. The gallery covers multiple themes but is compositionally imbalanced: modern romance and marriage/family melodrama alone account for 42.9% of titles, and the three modern-urban groups together account for 82.9%. At the shot-record level, the same modern romance/family group contributes 49.7% of the 101,763 annotated shots, whereas the wilderness-survival singleton contributes only 1.4%. We use "urban rise / identity reversal / supernatural" and "urban revenge / suspense / legal / business intrigue" as compact dataset-oriented labels; these labels summarize recurring plot motifs rather than official platform genres. The directory-to-table correspondence is checked against the released file inventory (35/35 names matched; no duplicate, missing, or extra title). This check validates inventory bookkeeping only and does not validate the semantic correctness of the thematic coding.

Table 16: Geometry stress in the learned alignment space. Gallery Recall@10 for modality removal is reported in Table 3 (full model 0.0954, without depth 0.0013, without pose 0.0184) and is not repeated here.

Control	Positive cos.	Angle agree.
Depth+pose (matched)	0.675	79.9%
Depth only	0.666	79.6%
Pose only	0.465	75.5%
Same-drama pose shuffle	0.666	80.3%
Global pose shuffle	0.664	80.8%
Both modalities, $\sigma = 0.25$	0.255	73.0%
4-level quantization	0.486	78.1%

Table 17: Primary-theme coding of the 35 DramaBoard titles. Each title is counted once.

Primary theme	Titles	Share
Modern romance / marriage and family	15	42.9%
Urban rise / identity reversal / supernatural	8	22.9%
Urban revenge / suspense / legal / business intrigue	6	17.1%
Costume / court	3	8.6%
Time travel / return to a past period	2	5.7%
Wilderness survival / adventure	1	2.9%

Structural hard cases and descriptor noise. The low-level hard-case slices identify where the structural prior is least reliable. Among 101,334 valid fused descriptors, the full-gallery normalized cluster entropy is 0.9558, whereas the depth-flat tail (5,067 records) has entropy 0.6541 and a 52.79% top-five-cluster share. The pose-missing proxy (4,274 records) is more concentrated, with entropy 0.3736 and a 91.81% top-five-cluster share; the pose-sparse tail (5,271 records) has entropy 0.4666 and a 78.11% top-five-cluster share. These slices are diagnostics of the released control JPEGs, not generated-image failure rates. Under standardized descriptor noise, cluster assignments remain stable for the fused representation (99.995% at $\sigma = 0.1$ and 99.92% at $\sigma = 2$), but degrade more for depth (99.97% to 91.29%) and pose (99.96% to 96.03%). The curves use a deterministic 200-shot sample. No downstream image, video, or human geometry conclusion is drawn from them.

To make noise robustness explicit beyond a single perturbation level, we also run a three-seed extension with 5,000 fixed queries and the same episode-level leakage prevention. The matched full condition has mean top-1 cosine 0.7173 ± 0.0048. Joint standardized Gaussian noise degrades it to 0.3446 ± 0.0007 at $\sigma = 0.25$, 0.2845 ± 0.0007 at $\sigma = 0.5$, and 0.2681 ± 0.0005 at $\sigma = 1.0$. Four-level descriptor quantization reduces the mean to 0.6575 ± 0.0038; 50% random pose dropout gives 0.6620 ± 0.0048, pose keypoint jitter at strength 0.1 gives 0.6608 ± 0.0051, and a pose distractor at strength 0.2 gives 0.7113 ± 0.0058. Depth blur with kernel 5, depth local occlusion with block 4, and a doubled people-count proxy give 0.6925 ± 0.0211, 0.6910 ± 0.0057, and 0.6895 ± 0.0056, respectively. These remain low-level descriptor curves rather than generated-image or learned retrieval results (Figure 10).

We additionally evaluate gallery construction using the released depth–pose controls. These diagnostics use low-level JPEG descriptors and report structural stress under controlled gallery settings. The full candidate set contains 101,334 depth–pose records. All query episodes are excluded from their candidate gallery. With 5,000 fixed queries, the original 100% gallery gives a mean top-1 cosine of 0.7118 and a 95.38% unique-reference ratio. Reducing the gallery to 10%, 25%, and 50% lowers mean similarity to 0.6521, 0.6762, and 0.6939, respectively. A cluster-balanced gallery (square-root allocation by the 128 fused descriptor clusters) increases normalized cluster entropy from approximately 0.956 for the original gallery to 0.985 at 10% and 0.981 at 50%, with only small additional similarity loss. Drama-balanced sampling drives the drama-level Gini coefficient close to zero but does not materially change similarity, indicating that drama imbalance and geometry-template concentration are separable factors.

As an independent seed-2030 recheck, the same gallery-size and balancing protocol gives a full-gallery mean top-1 cosine of 0.7110. At 10% gallery size, random and drama-balanced sampling give 0.6525 and 0.6526, respectively, whereas square-root cluster balancing gives 0.6535 and raises the

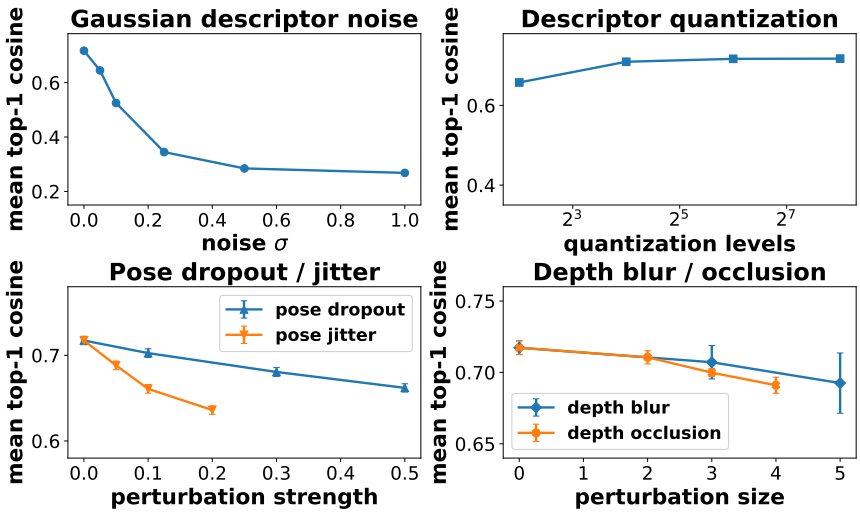

Figure 10: Extended three-seed low-level retrieval degradation curves for Gaussian noise, quantization, pose dropout/jitter, and depth blur/occlusion. Mean top-1 cosine is shown with sample standard deviation; these are descriptor-level controls.

Table 18: Gallery-size and balancing stress test (5,000 queries).

Gallery	Size	Mean sim.	Cluster entropy	Drama Gini
Original	100%	0.7118	0.9558	0.1402
Random	10%	0.6521	0.9567	0.1473
Drama-balanced	10%	0.6512	0.9567	0.0009
Cluster-balanced	10%	0.6520	0.9847	0.1425
Random	25%	0.6762	0.9544	0.1413
Drama-balanced	25%	0.6757	0.9560	0.0002
Cluster-balanced	25%	0.6765	0.9845	0.1479
Random	50%	0.6939	0.9557	0.1377
Drama-balanced	50%	0.6938	0.9559	0.0009
Cluster-balanced	50%	0.6942	0.9812	0.1453

selected-gallery normalized cluster entropy to 0.9881 (retrieved entropy 0.9743). The 25% and 50% cluster-balanced variants give mean similarities 0.6775 and 0.6940, with selected-gallery entropies 0.9878 and 0.9857. This recheck supports the direction of the structural balancing diagnostic, but it is not a trained-policy or downstream generation ablation.

Hard structural subsets On the same 10,000-query stress sample, environment-only shots and controls with a missing-pose proxy show substantially more concentrated retrieved clusters than the full sample. The latter is the clearest failure mode: only 12 of 128 fused clusters are retrieved and the top five account for 89.60% of references. Three-or-more-person and over-the-shoulder shots remain less concentrated, but their mean top-1 similarity is below or close to the overall 0.7130 value. These results identify where learned-embedding and generated-image tests should concentrate their evaluation budget. To make that follow-up budget reproducible, we freeze a 200-shot structural hard-case manifest with 25 disjoint shots from each of eight categories: environment-only, extreme close-up, high angle, low angle, over-the-shoulder, missing-pose proxy, three-or-more-person, and wide/long shots. The manifest is generated deterministically with seed 2026 and reproduces exactly across copies of the fixed data release. It defines the sampling protocol only; no generation-level score or human geometry judgment is inferred from membership in this manifest.

The table reports top-five individual-reference and top-five-cluster shares as separate concentration units; neither statistic implies that a learned policy or downstream generator has collapsed.

Table 19: Hard-subset structural retrieval diagnostic.

Subset	Queries	Mean sim.	Clusters	Top-5 cluster share
All	10,000	0.7130	127	10.78%
Environment-only	309	0.8725	77	49.19%
Three or more people	397	0.7100	98	23.93%
Over-the-shoulder	341	0.6934	93	20.82%
High angle	449	0.7329	107	27.39%
Wide/long shot	755	0.7642	115	25.70%
Missing-pose proxy	423	0.9576	12	89.60%

Table 20: Concentration diagnostics selected from the reference-seed, 10,000-query transparent descriptor stress test.

Subset	Queries	Unique ref.	Top-5 ref.	Eff. clusters	Top-5 cluster
All matched	10,000	91.26%	0.26%	104.23	10.78%
Environment-only	309	94.82%	5.18%	29.40	49.19%
High angle	449	97.55%	2.67%	65.38	27.39%
Over-the-shoulder	341	98.53%	2.93%	68.30	20.82%
Three or more people	397	97.98%	2.52%	65.66	23.93%
Wide/long shot	755	97.35%	1.72%	68.61	25.70%
Missing-pose proxy	423	89.60%	5.67%	6.14	89.60%

A.16 Prompt Templates

The templates below cover the annotation, training, generation, and evaluation stages of DramaDirector. Listings 1 and 2 specify episode-level semantic annotation and shot-level transcript correction. Listings 3 and 4 provide the storyboard annotation template and a continuation example for SFT. Listing 5 supplies the depth–pose conditioning instructions appended during first-frame generation.

Listings 6, 7, and 8 give the image, video, and human evaluation rubrics, respectively.

Listing 1: Episode-level semantic annotation prompt.

```
Episode-level Semantic Annotation Prompt
You are a professional AI drama analyst and a subtext reasoning expert. I will provide
    you with a dynamic video. Please analyze the video and output a character list, a
    plot summary, and a global dialogue list...
Requirements: Output must be entirely in simplified Chinese. Output must be pure JSON
    only. Do not include explanations or extra text.
Fields to extract:
1. characters: Identify all characters and provide Chinese names and objective
    appearance descriptions. If a character cannot be recognized, return [].
2. plot_summary: Summarize the core events of the episode.... dialogues:
3. Extract the dialogue in the video: Do not omit true spoken lines. Do not fabricate
    dialogue if none is...
```

Listing 2: Shot-level transcript correction prompt.

```
Shot-level Transcript Correction Prompt
You are a professional screenplay transcript correction assistant. Below is the
    information for one short-drama episode...
Correction Rules:
1. Use the dialogue reference to fix homophone errors, wrong names, and semantically
    incorrect ASR outputs.
2. If a transcript is an empty string, keep it empty. Do not add any content.
3. Keep each line concise. Do not expand or rewrite beyond correction.
4. Preserve all JSON keys exactly. The number and order of shot indices must remain
    unchanged.
Output Constraints: Output the corrected JSON object only...
```

Listing 3: Shot-level storyboard annotation prompt.

```
You are a professional storyboard designer. I will provide a sequence of video frames,
    each optionally paired with an ASR transcript, together with global information
    such as the character list, plot summary, and dialogue reference. Please generate
    one structured storyboard object for each frame.
Overall Requirements:
1. Be objective, precise, and visually grounded.
2. Emphasize fine-grained body movements and facial expressions.
3. Do not invent unverifiable mental states.
Hard Constraints:
1. One-to-one: if given N frames, you must output N JSON objects.
2. If speech spans multiple frames, split it into meaningful parts.
3. Do not fabricate dialogue. If silent, set dialogue, speaker, and emotion to null.
4. Do not use literary or psychological wording such as "seems", "as if", or "suggests".
5. Subtitle text must only appear in the dialogue field, not in descriptive fields.
Field Specifications:
- shot_scale: choose from [extreme close-up / close-up / medium close-up / medium / long
     shot / ...]
- camera_angle: choose from [high angle / low angle / eye level / over-the-shoulder /
    ...]
- camera_motion: choose from [static / push-in / pull-back / pan / tracking / ...]
- subjects: for each visible character, provide {name, gender, clothing, position,
    action, expression}
- background: scene type + major props + lighting / color tone
- description_narrative: a fluent shot-level narrative description grounded in visible
    content
- Dialogue, speaker, emotion, duration
Output Constraints:
- Output JSON array only. Each object must contain exactly: index, shot_scale,
    camera_angle, camera_motion, subjects, background, description_narrative, dialogue,
     speaker, emotion, duration
```

Listing 4: Example SFT prompt for storyboard continuation. Long contexts are truncated for readability.

```
System: You are a professional film and television storyboard scriptwriting assistant.
    Based on the provided plot summary, character information, and existing storyboard
    descriptions, continue generating the next batch of storyboard descriptions.
[Output Format] Strictly output a JSON array, where each element is a storyboard object.
     The fields should follow this template: [ { "index", "shot_scale": "Shot scale", "
    camera_angle": "Camera angle", "camera_motion": "Camera motion", "subjects": [{"
    name": "Character name", "gender": "Gender", "clothing": "Clothing description", "
    position": "Position in the frame", "action": "Action description", "expression": "
    Expression description"}], "background": "Environment description", "
    description_narrative": "What happens in the shot (narrative prose)", "dialogue": "
    Original dialogue text", "speaker": "Name of the speaking character", "emotion": "
    Emotion tag", "duration" }
[Notes]
index starts from 0 and should be numbered consecutively with the existing storyboards.
    duration unit is in seconds, kept to one decimal place. Do not output any extra
    text other than the JSON array itself. When all storyboards for this episode are
    completely generated, start a new line after the right bracket at the end of the
    JSON array and output
User: [Plot Summary] Mu Zhenguo flies into a rage over Mu Xingran's private life issues
    ...
[Characters] Mu Xingran: A young woman with long hair wearing a red slip dress with a
    white cardigan...
[Generated Storyboard Descriptions] "index", "shot_scale": "Medium close-up", "
    camera_angle": "Eye-level", "camera_motion": "Static", "subjects": [{"name", "
    gender": "Female"... Please continue generating the subsequent storyboard
    descriptions...
Assistant: "index", "shot_scale": "Medium close-up", "camera_angle": "Eye-level", "
    camera_motion": "Static"... "description_narrative": "Medium close-up static shot,
    Mu Wanqing stands sideways, looking to the right, with a surprised expression..."
```

Listing 5: Conditioning instructions appended to the storyboard-derived first-frame prompt.

```
Character Reference: Use Image as the character reference.
Spatial & Depth: Use the depth map in Image {depth_idx} as a spatial structure reference
    , prioritizing it to guide scene hierarchy, perspective relationships, and the
    overall spatial layout.
Pose Requirements (Environment Only): Use the skeleton map in Image {pose_idx} only as a
     light compositional reference. Do not generate any characters, body parts, or
    humanoid silhouettes based on it. This is a pure environment...
Pose & Composition Requirements (With Characters): Use the skeleton map in Image {
    pose_idx} as a character pose reference, prioritizing it to guide the general
    actions and limb orientation of the characters. Character movements can reference
    the skeleton map but must not rely entirely on it...
Character movements must follow natural physical laws. Poses should be natural and
    appropriate, with coordinated limb relationships. Avoid overly stiff, exaggerated,
    or unnatural poses. If the depth map causes distorted character ...
General & Style Requirements: Do not include any text, subtitles, dialogue bubbles, or
    other readable text in the image.
```

Listing 6: Detailed evaluation prompt and rubric for static image generation quality.

```
You are a strict evaluator for generated IMAGE quality. Your rubric is designed for
    either a VLM judge or a trained human judge. You will receive: Task prompt: {
    task_prompt}; Candidate image to evaluate: {candidate_image}
Goal: Evaluate the candidate image from six dimensions only: Aesthetics, consistency,
    faithfulness, attribute_binding, position, action
Important:
1. This is not a pixel-similarity task.
2. Judge only what is visually supported. Do not assume a hidden or blurry detail is
    correct.
3. Be strict on prompt-critical constraints and generous only when the evidence is
    clearly visible.
4. For image evaluation, "action" refers to STATIC action expression: whether a single
    frame captures a convincing, prompt-aligned decisive moment. Do not score temporal
    continuity, because there is no time axis.
Scoring scale: 1 = fully consistent with the metric definition; the evidence is clear
    and convincing. 0.5 = partially consistent; the intent is visible but incomplete,
    ambiguous, or only partly correct. 0 = inconsistent, missing, contradictory, or not
     verifiable.
Metric definitions:
1. Aesthetics: Judge the image's photographic, cinematic, and first-frame visual quality
     under a strict realistic standard. Prefer images that are suitable as stable first
     frames for realistic video generation: clear subject, natural lighting, coherent
    color, believable depth, stable composition, and grounded scene geometry...
2. Consistency: Judge whether the image is internally consistent and consistent with the
     prompt's stable entity-level constraints. Consider whether the main subjects,
    objects, materials, identities, and scene elements are presented without
    contradiction, and whether stable traits remain coherent within the single image...
3. Faithfulness: Judge whether the image satisfies the core semantic content of the text
     prompt...
4. attribute_binding: Judge whether prompt-specified attributes are bound to the correct
     subjects or objects...
5. Position: Judge whether the image satisfies the positional and spatial constraints in
     the text prompt...
6. Action: Judge the static action expression of the image...
```

Listing 7: Detailed evaluation prompt and rubric for temporal and dynamic video generation quality.

```
You are a strict evaluator for short-drama continuation storyboards and generated videos.

Inputs: plot context {plot_context}; continued storyboard {storyboard_continuation};
    generated video as the visual input.
Score only the judge metrics below. Similarity and DIS are automatic and are not scored
    here.
A. Storyboard narrative quality (text only; do not use the video): Relevance, Coherence,
    Engagement
B. Instruction following: Faithfulness, Purity
C. Video intrinsic quality: Aesthetics, Consistency
Scale: 1 = clear, complete, and convincing; 0.5 = partial, ambiguous, or incomplete; 0 =
    missing, contradictory, or not verifiable.
1. Relevance: the continuation stays grounded in the given plot context.
2. Coherence: the shot sequence is internally consistent in causality and pacing.
3. Engagement: the continuation carries short-drama dramatic interest.
4. Faithfulness: the video realizes the core content of the continued storyboard.
5. Purity: the video adds no prominent element that the storyboard does not state or
    reasonably imply.
6. Aesthetics: the video has clear subjects, natural lighting, and stable cinematic
    composition.
7. Consistency: identity, wardrobe, scene, and motion stay stable over time, independent
    of the text.
```

Listing 8: Shot-level human evaluation rubric.

```
Goal: Evaluate the candidate shot from six dimensions only: Relevance, Engagement,
    Faithfulness, Purity, Aesthetics, Consistency.
Important:
1. This is not a loose similarity task.
2. Judge only what is clearly visible in the shot.
3. Be strict on prompt-critical constraints: identity, wardrobe, key action, scene type,
    explicit details.
4. If the prompt requires a concrete action, the shot must show the action happening,
    not just a pose.
5. For Consistency, judge stability within the shot itself...
Scoring scale:
5 = fully consistent with the metric definition; evidence is clear and convincing
4 = mostly consistent; minor weakness
3 = partially consistent; clear but non-fatal problems
2 = weakly consistent; major problems
1 = inconsistent, missing, contradictory, or not verifiable
Metric definitions:
1. Relevance: Judge whether the shot prompt is well grounded in the established plot
    context, character identities, relationships, emotions, and scene state.
2. Engagement: Judge whether the shot contributes meaningful dramatic, emotional, or
    informational value, rather than feeling flat, generic, or filler-like.
3. Faithfulness: Judge whether the video satisfies the core semantic content of the shot
    prompt, especially the key action, event, emotional beat, and subject relation.
4. Purity: Judge whether the video avoids important unprompted elements such as extra
    people, wrong environment, intrusive captions, or clearly wrong reused assets.
5. Aesthetics: Judge the visual and cinematic quality of the shot under a strict
    realistic standard: subject clarity, lighting, composition, depth, texture, and
    overall polish.
6. Consistency: Judge whether the shot remains internally stable over time, including
    subject identity, clothing, background, object continuity, and motion progression.
```

A.17 SPECIFICATION LISTINGS

Listing 9: Implemented reward specification and abnormal-output handling.

```
R_ret | appearance-reduced visual queries; shot to completion | [0,1]
  Mean top-k cosine over shots, contrasted with one sampled negative episode; k=1.
  Negative episode: another episode of the same drama, else another drama.
  No-key debug mode returns 0.5 for a missing target, empty query, or unset key; not a
       reportable result.
  With an API key, a failed embedding request raises a failed-run exception; malformed
       counts, dimensions, or non-finite vectors are rejected.

R_vid | complete ordered narrative descriptions | [0,1]
  One equal-weight judge score over five dimensions, clipped to [0,1] and cached by the
       full sequence.
  After three API or JSON failures, the affected batch returns 0.5 and the fallback is
       logged.

R_fmt tilde | parsed JSON shots; field and subject structure | [-1,1]
  2*R_fmt-1, with 60/30/10 weights for key presence, non-empty content, and subject
       subfields.
  Does not validate camera enums, duration ranges, or index continuity.

Total R | one completion | [-0.1,1.3]
  alpha_s*R_ret + beta_s*R_vid + gamma*R_fmt_tilde.
  At the start of annealing a completion lies in [-0.1,1.3]; after alpha_s reaches 0.5
       the range is [-0.1,1.1].
  Truncated JSON keeps complete objects decoded before the cut; repairs are logged.
```

Listing 10: Recorded computational measurements and commercial API references.

```
Base Qwen3-8B diagnostic | 80 completions; 20 prompts x 4 seeds | load 4.88 s;
     generation 1365.39 s; peak GPU 18.70 GiB; p50/p95 35.86/36.18 s | 17.07 s/
     completion
Released-adapter diagnostic | same prompts, seeds, decoding | load 5.86 s; generation
     1524.66 s; peak GPU 19.68 GiB; p50/p95 43.44/48.29 s | 19.06 s/completion
Expanded released-model diagnostic | 20 prompts x 17 seeds; 680 completions | base/
     adapter generation 5800.24/6702.16 s; peak GPU 17.45/18.36 GiB | 17-seed diversity
     diagnostic
200-prompt four-seed stability | 200 prompts x 4 seeds; 800 completions/model | base/
     adapter generation 7471.71/12697.40 s; load 7.74/25.56 s; peak GPU 24.31/27.56 GiB
     | batch sizes base 8/8, adapter 8/2
200-prompt hard-case scale-up | 200 prompts x 4 seeds; 800 completions/model | base/
     adapter generation 8206.70/10176.40 s; load 11.22/13.72 s; peak GPU 19.71/21.42 GiB
      | planner hard-case diagnostic
Offline anti-collapse GRPO | 2 updates; 2 prompts; 2 generations/prompt | training 48.81
      s | reconstructed-SFT init; format + structure rewards
Format-only GRPO control | 2 updates; same prompts | training 49.17 s | diversity weight
      0
Offline memory probe | 1 update; same init | training 25.53 s; peak GPU 45.1 GiB | one-
     step probe
20-prompt offline GRPO | 2 updates; 20 examples; 2 generations | structure/format
     training 39.46/37.16 s; peak GPU 46.6/15.4 GiB | 0% exact JSON, 80% non-empty
40-prompt offline GRPO | 4 updates; 40 examples; 2 generations | structure/format
     training 73.76/60.42 s; peak GPU 46.6/15.4 GiB | seed 2048
Matched proxy generation | 20 prompts; one completion | base/SFT generation
     210.92/268.77 s; load 3.35/5.02 s; peak GPU 19.8/21.5 GiB | seed 2048; planner only
Depth-control descriptors | 106,349 JPEGs | 86.9 s | low-level feature extraction
Pose-control descriptors | 106,349 JPEGs | 98.7 s | low-level feature extraction
Geometry noise retrieval | 3 seeds x 44 conditions x 5,000 queries | 18.9-19.8 s/seed |
     descriptor perturbation
Commercial APIs | released per-shot pipeline | see API cost and latency tables | public
     list prices
```