WANPE: TOWARDS CINEMATIC PROMPT ENHANCEMENT FOR MODERN TEXT-TO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation begins in text space by authoring a cinematic screenplay, then materializes into pixels. As contemporary video generators scale to 30 seconds and faithfully follow complex conditions, the textual prompt largely directs the production, planning how actions, camera trajectories, lighting, and sound unfold across multi-shot sequences. In this paper, we present **WanPE**, a 397B-parameter prompt enhancement model trained on 1.05M real-world videos to master director-level cinematic planning. WanPE formulates shot-level cinematic plans via video-grounded reverse construction, and employs Semantic-Consistency GRPO (SC-GRPO) to faithfully preserve user requirements across shots and over time. To benchmark this capability, we curate **WanPEval**, a human-annotated testbed covering durations from 5 to 30 seconds across varying intent granularities, supported by ~ 11K blind pairwise assessments. When powering Wan3.0's video generator, WanPE-397B boosts human preference over raw user prompts by 10.66-18.84 points at 5-15 seconds, and by a dramatic 50.86 points in the 30-second arena. Ablation studies show that reverse construction demonstrates clear superiority over forward rewriting, while SC-GRPO robustly preserves semantic fidelity across model scales. Ultimately, WanPE leads all evaluated commercial offerings at 5-15 seconds and remains competitive with Seedance 2.5 at 30 seconds. The demo is available here.

1 INTRODUCTION

Recent advances in text-to-video (T2V) generation have enabled systems such as Wan3.0 (Wan Team, 2026) and Seedance 2.5 (ByteDance Seed, 2026) to generate up to 30 seconds of cinematic-quality video, with expressive camera movements, realistic lighting, and coherent narratives. These systems rely on two tightly coupled components: a prompt enhancer that transforms user input into textual conditions, and a video generator that renders these conditions into video. As video generators process longer contexts and faithfully follow increasingly complex instructions, textual control now reaches the level of cinematic direction. Accordingly, the role of prompt enhancement shifts from enriching user requests with descriptive details to directing the entire production, planning how actions, camera, lighting, dialogue, and sound unfold coherently across shots and over time.

Many recent works have explored prompt enhancement for T2V generation, including learned prompt rewriting from user requests to detailed descriptions and multi-step LLM refinement (Wang et al., 2025; Cheng et al., 2025; Ji et al., 2025; Wu et al., 2026; Jia et al., 2026; Xue et al., 2025; Ma et al., 2024; Yang et al., 2026; Gao et al., 2026; Long et al., 2026). Historically, constrained by bounded context and weak instruction following in early video models, such as Wan2.2 (Wan Team et al., 2025), prompt enhancement merely served as cosmetic visual enrichment. Today, modern video generators have broken these bottlenecks, stretching horizons to 30 seconds and unlocking unprecedented headroom for expressive textual control. In this paper, we break away from the conventional paradigm of descriptive prompt rewriting. Instead, we conceptualize text as the living flow of the video itself, the very blueprint where temporal cadence, camera choreography, and multi-shot transitions unfold before pixel rendering. Guided by this philosophy, we fundamentally redesign the prompt enhancement framework into a cinematic planning architecture. Realizing this architecture introduces two key challenges: learning coherent cinematic planning across interdependent production components, and preserving user-specified requirements throughout the unfolding plan.

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.



Figure 1: **Left:** Earlier models use optional prompt expansions, while modern video generator can process longer contexts and follow complex instructions. **Right:** t-SNE shows WanPE outputs align closely with video-grounded captions, while forward-based enhancement exhibits a distribution gap.

For the first challenge, most T2V prompt enhancers are built around forward expansion from user requests to richer conditions. Their outputs follow a synthetic rewriting distribution. This creates a training–inference mismatch for downstream video generators trained on video-grounded captions. As prompt enhancement scales toward coherent cinematic planning, this conditioning mismatch reflects an asymmetry beyond linguistic style, as illustrated in Figure 1. For example, to expand a request for a two-person action scene, a forward enhancer must not only imagine a sequence of fast interactions, but also determine how camera movement, lighting, audio, and other elements evolve with these interactions over time. However, a professionally filmed action sequence embodies coordination across acting, choreography, cinematography, editing, lighting, and sound. These real-world videos provide realized cinematic structure and video-grounded distribution targets.

For the second challenge, coherent cinematic planning elevates semantic preservation into a global consistency requirement over the entire plan. The enhancer must faithfully propagate user-specified subjects, actions, dialogue, camera instructions, and event order to every relevant shot while preserving their bindings and temporal relationships across the full sequence. At the same time, newly introduced details must remain compatible with both the original request and preceding planning decisions. As the plan evolves shot by shot, requirements may otherwise be omitted, altered, assigned to the wrong subject or shot, or contradicted by later planning decisions.

In this paper, we introduce **WanPE**, a prompt enhancer trained on 1.05M real-world videos for director-level cinematic planning. WanPE reverses conventional supervision construction by deriving a hierarchical cinematic condition y from each high-quality video and reconstructing a compatible user request x. To preserve user requirements throughout the cinematic plan, we optimize the enhancer with SC-GRPO using a nine-dimensional reward that penalizes omissions, alterations, incorrect bindings, and temporal inconsistencies. We evaluate 4B–397B variants on **WanPEval**, a human-annotated 5-30-second testbed with varying intent granularities, using text-level scoring and $\sim$11K blind pairwise video assessments. WanPE-397B improves over raw prompts by up to 18.84 points at 5-15 seconds and 50.86 points at 30 seconds. It leads evaluated commercial offerings at 5-15 seconds while remaining competitive with Seedance 2.5 at 30 seconds. Ablations show reverse construction outperforms forward rewriting, while SC-GRPO improves semantic consistency by 18.6-23.3 points across scales. WanPE transfers across video generators after format adaptation.

Our contributions are summarized as follows:

- We redesign prompt enhancement as a scalable cinematic planning architecture, transforming text into the blueprint that orchestrates actions, camera choreography, audio, and narrative across shots and through time before pixel rendering.
- We realize this architecture by learning realized cinematic structure from real-world videos through video-grounded reverse SFT and faithfully preserving all user requirements throughout the unfolding cinematic plan through SC-GRPO.
- We introduce WanPEval, spanning diverse request complexities and generation durations from 5 to 30 seconds, and scale WanPE from 4B to 397B to demonstrate consistent effectiveness across model scales and downstream video generators.

2 METHODOLOGY

In this section, we present how WanPE realizes director-level cinematic planning while preserving user requirements across shots and over time. WanPE first formalizes these dual objectives (Section 2.1), learns coherent cinematic planning through video-grounded reverse construction (Section 2.2), and strengthens semantic fidelity with SC-GRPO (Section 2.3). Finally, WanPEval enables text and video-level evaluation across request granularities and generation durations (Section 2.4).

2.1 PROBLEM OVERVIEW AND FORMULATION

A modern text-to-video (T2V) system composes a prompt enhancer π_θ with a video generator G_ϕ. Given a natural-language user request x from the user-prompt distribution p_{u}, the prompt enhancer π_θ produces a textual condition y, which guides the video generator G_ϕ to synthesize a video $\hat{v}$:

$$x \sim p_{\mathrm{u}}, \qquad y \sim \pi_\theta\left(\cdot \mid x\right), \qquad \hat{v} \sim G_\phi\left(\cdot \mid y\right). \tag{1}$$

The textual condition y should preserve all requirements in the user request x. For a given x, multiple outputs may satisfy these requirements. We denote the set of such outputs by $\mathcal{Y}_{\mathrm{sem}}\left(x\right)$:

$$\mathcal{Y}_{\mathrm{sem}}\left(x\right) = \left\{y \mid c\left(y\right) = 1,\ \forall c \in \mathcal{C}\left(x\right)\right\}, \tag{2}$$

where $\mathcal{C}\left(x\right)$ denotes the set of user-specified semantic and instructional constraints in x, and $c\left(y\right) = 1$ indicates that y satisfies the requirement c without omission, alteration, or contradiction.

The output distribution of the enhancer should align with the conditioning distribution of the video generator G_ϕ. The video generator G_ϕ is trained on real-world videos paired with video-grounded captions. We denote the video-grounded caption distribution by $p_{\mathrm{vg}}\left(y\right)$.

Combining semantic preservation with distributional alignment, we define the ideal target distribution as the video-grounded caption distribution restricted to semantically valid outputs:

$$p^*\left(y \mid x\right) = \frac{p_{\mathrm{vg}}\left(y\right)\mathbb{I}[y \in \mathcal{Y}_{\mathrm{sem}}\left(x\right)]}{Z\left(x\right)}, \tag{3}$$

where $Z\left(x\right)$ normalizes the distribution. We then learn π_θ to match p^* over $x \sim p_{\mathrm{u}}$:

$$\theta^* = \arg\min_\theta \mathbb{E}_{x\sim p_{\mathrm{u}}}\left[D_{\mathrm{KL}}\left(p^*\left(\cdot \mid x\right) \,\|\, \pi_\theta\left(\cdot \mid x\right)\right)\right]. \tag{4}$$

Modern video generators can synthesize videos of up to 30 seconds, process longer textual contexts, and follow more complex instructions. This conditioning capacity creates unprecedented room for textual control. To harness this capacity, we redesign the prompt enhancement framework into a cinematic planning architecture. Learning such conditions poses two challenges. **(i) Learning coherent cinematic planning.** π_θ must learn the joint organization of interdependent cinematic elements while matching p_{vg}. **(ii) Maintaining long-range semantic fidelity.** User-specified constraints such as identity, dialogue, and camera must stay consistent across temporally ordered shots.

2.2 VIDEO-GROUNDED SUPERVISED FINE-TUNING

To learn how cinematic elements are jointly organized across shots and over time while aligning the enhancer's outputs with p_{vg}, we first perform supervised fine-tuning on prompt–caption pairs (x, y) satisfying $y \in \mathcal{Y}_{\mathrm{sem}}(x)$. Starting from x and enriching it along predefined descriptive dimensions can readily satisfy this semantic requirement. However, the added content and its cinematic organization are determined by the rewriting procedure, causing the resulting targets to follow a synthetic rewriting distribution rather than p_{vg}. Our key design is to reverse the pair-construction process: we first obtain y by captioning a real-world video and then derive a semantically compatible x from y.

Real-world video collection and filtering. To support alignment with p_{vg}, we collect a large and diverse set of real-world videos that are either publicly available or licensed for use. We segment long-form videos into clips of at most 30 seconds and apply multi-stage filters for technical validity, visual quality, and motion quality. After filtering, the resulting dataset contains $\sim 1.05\mathrm{M}$ clips spanning ten content dimensions with diverse camera movements. The details are in the Appendix A.

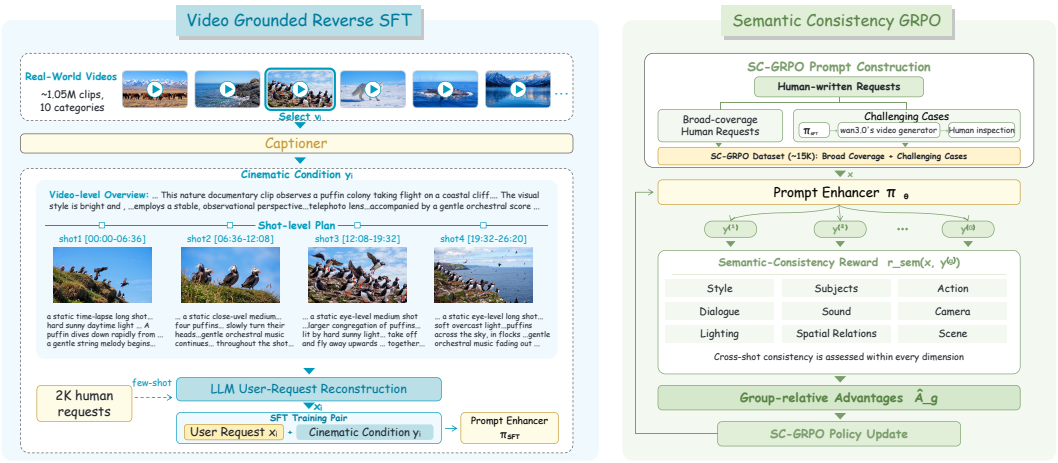


Figure 2: WanPE training pipeline. Video-grounded reverse SFT and semantic-consistency GRPO.

Video-grounded cinematic captioning. For each curated video clip v_i, we use a multimodal video captioner f_{cap} to analyze its visual and audio content and produce a textual cinematic target y_i:

$$y_i = f_{\text{cap}}(v_i; s_i), \tag{5}$$

where s_i denotes the category-specific captioning instruction for v_i.

The target y_i organizes v_i hierarchically into a video-level summary and temporally ordered shot-level descriptions with timestamps. Each shot specifies its composition, subjects, actions, lighting, camera movement, transitions, dialogue, music, and sound effects, together with their progression over time. The instruction s_i adapts this structure to the category of v_i. We keep captions that pass checks for structural completeness, timestamp validity, and consistency with the source video, and use them as video-grounded targets from p_{vg} for SFT. The details are in the Appendix B.

User-request reconstruction. In reverse pair construction, we derive a user request x_i from video-grounded target y_i. Summarizing y_i tends to retain the caption's structure and detail, producing a compressed caption rather than a natural user request. We therefore employ `gpt-5.4` with few-shot prompting, using a pool of 2K requests written by human annotators across ten content categories.

Let d_i denote the category of y_i, and let $\mathcal{H}_{d_i}$ denote the corresponding request pool. We sample five demonstrations $\mathcal{E}_i$ from $\mathcal{H}_{d_i}$. Given y_i and $\mathcal{E}_i$, the LLM reconstructs

$$x_i = f_{\text{LLM}}(y_i; \mathcal{E}_i). \tag{6}$$

The reconstruction prompt requires x_i to contain only requirements supported by y_i, ensuring $y_i \in \mathcal{Y}_{\text{sem}}(x_i)$. The demonstrations $\mathcal{E}_i$ guide x_i toward language and specificity of natural user requests.

Supervised fine-tuning. The reverse construction yields the SFT dataset $\mathcal{D}_{\text{SFT}} = \{(x_i, y_i)\}_{i=1}^{N}$. We fine-tune the prompt enhancer π_θ by minimizing the negative log-likelihood of each video-grounded cinematic condition given its reconstructed user request:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log \pi_\theta(y_i \mid x_i). \tag{7}$$

Through this objective, the enhancer learns to transform natural user requests into video-grounded cinematic conditions that capture the joint organization of elements across shots and over time.

2.3 Semantic-Consistency GRPO

To preserve user requirements as cinematic conditions unfold across shots and events, we introduce Semantic-Consistency GRPO (SC-GRPO). It applies Group Relative Policy Optimization (Shao et al., 2024) with a semantic-consistency reward penalizing omissions, alterations, inconsistent bindings between subjects, actions, and dialogue, and temporal inconsistencies across shots.

Training data construction. To cover diverse requests and challenging cases, human annotators write T2V prompts for videos up to 30 seconds. We generate videos with the Wan3.0 video generator, conditioned on outputs from the SFT enhancer. Through visual assessment, annotators flag prompts whose videos inadequately realize the requested content. We combine these challenging cases with human-written prompts to form $\mathcal{D}_{\text{SC-GRPO}}$, containing approximately 15K prompts.

Semantic consistency reward. We use `Qwen3.7-Max` to assess whether y preserves the specified constraints $\mathcal{C}(x)$, producing a text-only reward $r_{\text{sem}}(x, y) \in [0, 100]$, with higher scores indicating better requirement preservation. Evaluation spans nine dimensions: style, subjects, actions, dialogue, sound, camera, lighting, spatial relations, and scene. The evaluator checks for omitted, weakened, altered, or contradictory user requirements, incorrect subject–attribute and speaker–dialogue bindings, action and shot ordering errors, and cross-shot conflicts. Semantically equivalent paraphrases and compatible elaborations are accepted, while semantic discrepancies are penalized by severity. Details are in the Appendix C.

Policy optimization. We initialize π_θ from the frozen reference π_{SFT}. For each $x \in \mathcal{D}_{\text{SC-GRPO}}$, we sample G conditions $\{y^{(g)}\}_{g=1}^{G}$ from $\pi_{\theta_{\text{old}}}(\cdot \mid x)$ and standardize their rewards $r_{\text{sem}}(x, y^{(g)})$ into group-relative advantages $\hat{A}_g$. We maximize

$$\mathcal{J}_{\text{SC-GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{T_g}\sum_{t=1}^{T_g}\left(\min\left\{\rho_{g,t}\hat{A}_g, \text{clip}(\rho_{g,t}, 1-\epsilon, 1+\epsilon)\hat{A}_g\right\} - \beta\mathcal{K}_{g,t}\right)\right], \quad (8)$$

where T_g is output length, $\rho_{g,t}$ the current/old policy ratio, and $\mathcal{K}_{g,t}$ the SFT-reference KL penalty.

2.4 WanPEval: Testbed and Evaluation

Evaluating modern T2V prompt enhancers requires diverse generation durations and request complexities. We therefore introduce **WanPEval**, a human-annotated testbed spanning 5 to 30 seconds and requests ranging from concise high-level intents to detailed shot-level instructions.

Data construction and curation. Human annotators write practical T2V requests across diverse topics, stratified by duration and granularity, from high-level intents to structured shot-level instructions. We review requests for semantic clarity, internal consistency, and temporal feasibility, remove duplicates, and exclude overlap with SFT and SC-GRPO training data. The curated WanPEval contains 249 requests. Details are in the Appendix E.

Text-level semantic evaluation. To evaluate semantic fidelity at text level, we apply the semantic consistency reward $r_{\text{sem}}(x, y)$ defined in Section 2.3 to each condition generated for WanPEval.

Expert video evaluation. The practical value of a prompt enhancer lies in the quality of downstream videos conditioned on its outputs, which requires jointly assessing request adherence, perceptual quality, and cinematic coherence. As automated metrics struggle to reliably capture these aspects, we adopt anonymous pairwise human preference evaluation following the Artificial Analysis Video Arena (Artificial Analysis), a widely recognized leaderboard for modern video generation models.

Each of the M methods generates one video per request, yielding up to $N\binom{M}{2}$ pairs across N requests. Evaluation involves 60 experts in screenwriting, directing, cinematography, and related film disciplines. Each pair is shown with request x, with identities hidden and left–right order randomized. Experts choose: A preferred, B preferred, both good, or both poor. See Appendix E.3 for the interface. For method k, let N_k denote its number of valid comparisons after excluding unsuccessful generations, and let N_{win}^k and N_{bg}^k denote its win and both-good counts, respectively. We compute its lower-bound, upper-bound, and overall preference scores as

$$S_{\text{lower}}^k = \frac{N_{\text{win}}^k}{N_k}, \qquad S_{\text{upper}}^k = \frac{N_{\text{win}}^k + N_{\text{bg}}^k}{N_k}, \qquad S^k = \frac{N_{\text{win}}^k + 0.5N_{\text{bg}}^k}{N_k}.$$

We additionally fit a Bradley–Terry model (Bradley & Terry, 1952) to account for opponent strength, treating both-good and both-poor outcomes as ties:

$$p_{kj} = \sigma(b_k - b_j), \qquad \sum_k b_k = 0, \qquad \text{BT}^k = 100\,\sigma(b_k).$$

Here, b_k denotes strength. We report BT^k as the preference score against a mean-strength opponent.

Table 1: Expert preference score and Bradley–Terry score on the 5–15 seconds subset of WanPEval.

Method	5 seconds		10 seconds		15 seconds		Overall	
	S	BT	S	BT	S	BT	S	BT
LTX-2.5	18.69	31.23	18.47	34.74	16.10	32.20	17.30	33.02
Kling 3.0	31.11	43.02	23.47	39.84	16.10	34.07	20.80	37.40
HappyHorse 1.1	20.28	32.82	24.00	40.72	26.80	42.39	24.89	40.46
MiniMax-H3	29.09	41.24	36.46	49.59	38.30	51.13	36.40	49.27
Seedance 2.0	46.20	54.20	41.40	52.16	44.31	56.90	43.42	54.70
downstream generator: Wan3.0's video generator								
Original request	46.08	57.13	33.63	49.40	30.59	47.69	33.80	49.59
+WanPE-4B	48.17	57.28	41.27	54.58	44.00	57.10	43.60	56.21
+WanPE-9B	50.91	59.76	44.40	57.06	45.78	58.24	46.00	58.01
+WanPE-35B	50.69	59.32	48.76	60.22	46.45	60.27	47.85	60.09
+WanPE-397B	**56.74**	**65.00**	**49.91**	**62.00**	**49.43**	**60.95**	**50.61**	**61.85**

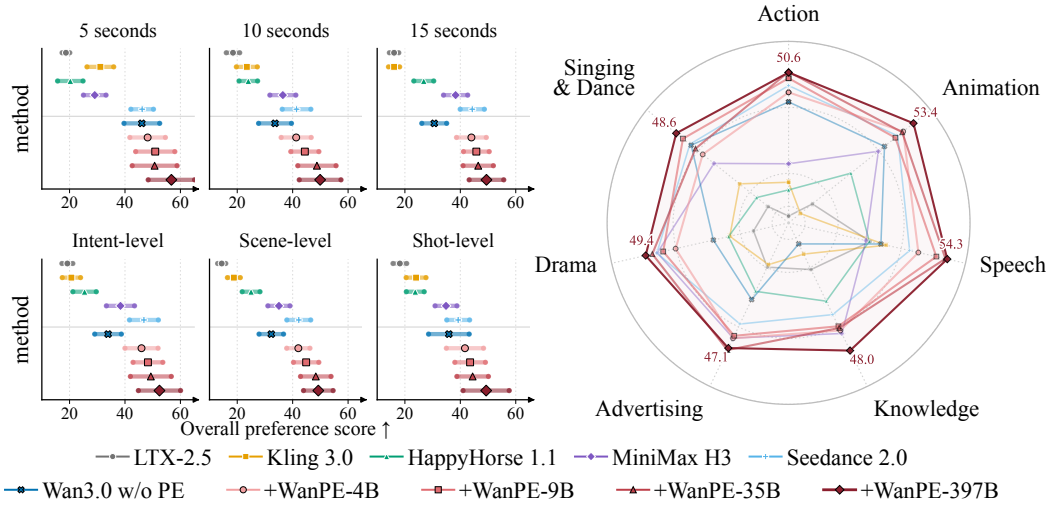


Figure 3: Fine-grained expert evaluation on the 5-15 seconds subset of WanPEval. **Left:** Preference scores across generation durations and request-granularity levels. Markers denote S, and horizontal intervals span $[S_{\text{lower}}, S_{\text{upper}}]$. **Right:** Preference scores S across seven content categories.

3 Experiments

In this section, we comprehensively evaluate WanPE from four perspectives: generation quality (Section 3.1), the effectiveness of reverse-constructed supervision (Section 3.2), transferability across downstream video generators (Section 3.3), and training analysis (Section 3.4).

Settings. We instantiate four model variants initialized from Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-35B-A3B, and Qwen3.5-397B-A17B (Qwen Team, 2026), denoted as WanPE-4B, WanPE-9B, WanPE-35B, and WanPE-397B, respectively. The experiments are conducted on 512 GPUs. Unless otherwise specified, each request is first enhanced by WanPE, and the output is then provided to Wan3.0's video generator (Wan Team, 2026) for video generation and evaluation. For an evaluation involving n methods, we construct all $\binom{n}{2}$ pairwise video comparisons for each request and aggregate the resulting judgments to compute the final scores. Details are provided in the Appendix F.

3.1 Overall Comparison

We evaluate the generation quality of WanPE on WanPEval. We compare our prompt-enhanced pipeline with Seedance 2.5 (ByteDance Seed, 2026), Seedance 2.0 (Seedance et al., 2026), HappyHorse 1.1 (HappyHorse Team, 2026), Kling 3.0 (Kling AI, 2026), MiniMax-H3 (MiniMax Team, 2026), and LTX-2.5 (Lightricks, 2026), and an original-request baseline without prompt enhancement. Given each system's maximum supported duration, we evaluate Seedance 2.5 against WanPE-397B on the 30-second subset and all other systems on the 5–15-second subset of WanPEval.

Table 2: Preference scores S on WanPEval. (i) 30-second subset against Seedance 2.5, under Wan3.0. (ii) Reverse-constructed vs. forward-based enhancement, under Wan3.0. (iii) Format-adapted WanPE vs. native enhancers on LTX-2.5 and MiniMax-H3 in two separate battles.

Method	Action	Anim.	Speech	Ad.	Sing & Dance	Drama	Know.	Overall
Exp1: 30-second subset								
Seedance 2.5	**68.18**	46.43	55.00	**81.25**	**60.00**	**56.00**	**60.71**	59.76
downstream generator: Wan3.0's video generator								
Original request	4.17	9.38	15.79	5.00	22.50	4.00	3.57	9.38
+WanPE-397B	50.00	**81.25**	**73.68**	75.00	47.50	53.85	57.14	**60.24**
Exp2: Reverse vs. forward enhancement								
downstream generator: Wan3.0's video generator								
Original Request	23.97	21.01	29.85	19.32	35.00	22.60	6.38	23.05
+ Forward Rewriting	39.57	38.43	42.13	43.57	38.61	34.62	41.89	39.49
+ Forward-target SFT	30.36	44.23	32.28	37.00	35.34	33.72	34.26	35.17
+ WanPE-397B-SFT	**46.48**	**55.98**	**48.31**	**49.32**	**45.51**	**51.54**	**51.23**	**49.86**
Exp3: Cross-generator transfer (5–15s)								
downstream generator: LTX-2.5-Base								
+ LTX-2.5-PE	8.33	18.33	30.00	**32.50**	7.50	26.67	25.00	21.11
+ WanPE-397B	**25.00**	**38.33**	**53.33**	27.50	**27.50**	**36.67**	**35.00**	**35.56**
downstream generator: MiniMax-H3-Base								
+ H3-Context-IR	**44.44**	31.03	33.33	39.47	37.50	29.31	**40.00**	35.92
+ WanPE-397B	29.63	**37.93**	**40.00**	**44.74**	**47.50**	**50.00**	**40.00**	**41.09**

Settings. For MiniMax-H3 and LTX-2.5, each request is first enhanced by the native prompt enhancer, H3-Context-IR or LTX-2.5-PE, and the enhanced condition is passed to the generator, H3-Base or LTX-2.5-Base. For the remaining baselines, the native prompt enhancer and video generator form an end-to-end API, so we submit each request directly and evaluate the returned video.

WanPE is strong and consistent on WanPEval. We first ablate PE scale under the same Wan3.0 video generator. Figure 3 shows that WanPE-397B ranks first on intent-, scene-, and shot-level requests, with scores of 52.44, 49.26, and 49.32. Compared with using no prompt enhancement, WanPE-397B delivers substantial gains of 18.55, 16.97, and 13.45 points across the three request granularities. We then compare WanPE-397B with other leading video generation systems. As shown in Table 1, it ranks first in both S and BT at every duration on the 5–15-second subset, with overall scores of 50.61 and 61.85, and outperforms Seedance 2.0 by 7.19 points in S. As shown in Table 2, it remains competitive with Seedance 2.5 on the 30-second subset, especially on animation and speech, where it scores 81.25 and 73.68. Details are provided in the Appendix G.

Prompt enhancement becomes increasingly important as the generation horizon grows. Under the same Wan3.0 video generator, WanPE-397B improves over the original-request baseline by 10.66, 16.28, and 18.84 points at 5, 10, and 15 seconds. On the 30-second subset, it raises S from 9.38 to 60.24. The gains therefore increase over 5–15 seconds and remain large at 30 seconds.

3.2 Reverse-Constructed Supervision vs. Forward Prompt Expansion

A central design choice of WanPE is to construct supervision in reverse from video-grounded captions. We isolate its effect through two forward-based baselines: (i) **Forward Rewriting.** 8 experts in computer science and film directing designed an instruction covering semantic fidelity, cinematic structure, temporal progression, camera language, lighting, narrative development, and audiovisual coherence. These dimensions match those of the video-grounded targets y in Section 2.2. On a separate validation set, Gemini 3.1 Pro (Google, 2026) rewrites the original requests and Wan3.0's DiT generates the videos. Following the blind protocol in Section 2.4, we refined the instruction over 27 versions. The final version is the Forward Rewriting baseline. (ii) **Forward-target SFT.** Using the reconstructed requests x from Section 2.2, we apply the same Forward Rewriting method to obtain targets y_{fwd}, and train Qwen3.5-397B-A17B on (x, y_{fwd}) with identical training settings.

Video-grounded learning provides stronger conditions than forward rewriting. As shown in Table 2, WanPE-397B-SFT attains an overall S of 49.86, outperforming Forward Rewriting by 10.37 points and Forward-target SFT by 14.69 points. It ranks first in all seven content categories. These results indicate that video-grounded cinematic captions provide more effective textual conditions than forward-constructed targets.

Table 3: (i) Semantic consistency evaluated by Gemini-3.1-Pro. † denotes results computed on the 5–15-second subset. (ii) Expert preference S for 397B SFT vs. WanPE-397B under Wan3.0.

(i) Semantic consistency

Method	Overall↑	5 seconds↑	10 seconds↑	15 seconds↑	30 seconds↑	Perfect↑	Failure↓
Fwd. Rewriting	90.7	93.5	91.8	90.3	89.4	55.0	10.0
LTX-2.5-PE	77.3	76.1	81.6	74.9	77.0	33.7	38.6
H3-Context-IR	88.3†	89.4	88.8	87.6	–	54.2†	17.3†
Ours							
WanPE-4B-SFT	66.5	67.0	68.4	66.7	64.2	20.1	56.6
+ SC-GRPO	85.3 +18.8	91.9 +24.9	83.1 +14.7	85.3 +18.6	84.8 +20.6	52.0 +31.9	27.4 −29.2
WanPE-9B-SFT	70.2	74.6	69.5	72.9	65.6	22.5	48.6
+ SC-GRPO	88.8 +18.6	94.0 +19.4	90.6 +21.1	87.9 +15.0	86.5 +20.9	53.8 +31.3	15.7 −32.9
WanPE-35B-SFT	73.1	69.0	70.1	77.3	71.7	21.3	42.6
+ SC-GRPO	96.4 +23.3	**99.3** +30.3	97.9 +27.8	95.0 +17.7	95.9 +24.2	80.7 +59.4	4.4 −38.2
WanPE-397B-SFT	75.5	73.2	70.0	81.5	73.4	29.7	36.9
+ SC-GRPO	**97.6** +22.1	99.1 +25.9	**98.0** +28.0	**96.9** +15.4	**97.6** +24.2	**85.5** +55.8	**2.8** −34.1

(ii) PE-scale expert preference on WanPEval

Method	Action	Anim.	Speech	Ad.	Sing & Dance	Drama	Know.	Overall
Video generator: Wan3.0's video generator								
Original Request	24.71	26.62	27.44	11.00	32.50	30.12	15.74	24.95
+ 397B SFT	47.02	47.37	40.24	54.00	37.50	38.75	34.26	42.70
+ WanPE-397B	**47.65**	**48.70**	**53.05**	**55.00**	**53.33**	**40.96**	**53.70**	**49.69**

3.3 Transferability of Cinematic Planning Across Video Generators

We test whether WanPE plans remain effective beyond Wan3.0's video generator. Because generators use different conditioning formats, we adapt WanPE to each native format while keeping the cinematic content. We use MiniMax-H3 and LTX-2.5, whose enhancers and generators are separately accessible. GPT-5.4 converts WanPE outputs into the formats of H3-Context-IR and LTX-2.5-PE, which are then fed to H3-Base and LTX-2.5-Base. Details are provided in the Appendix H.

WanPE transfers across video generators. After format adaptation, WanPE-397B outperforms the native enhancers on both generators (Table 2, Exp3), improving S by 14.45 points on LTX-2.5-Base and 5.17 points on MiniMax-H3-Base. These results suggest that its learned cinematic organization remains effective across generator-specific conditioning formats.

3.4 Semantic Consistency and SC-GRPO Analysis

In this subsection, we evaluate whether SC-GRPO improves semantic consistency and whether this improvement also benefits downstream video generation.

SC-GRPO consistently improves semantic consistency. Figure 4 shows that the mean training reward increases substantially across four model sizes, with larger models reaching higher final values. These gains are reflected on WanPEval. As shown in Table 3, SC-GRPO improves the Gemini-3.1-Pro overall score by 18.6-23.3 points across four model sizes, with consistent gains at every evaluated duration. It increases the proportion of perfect outputs and reduces failures. Additional results are provided in Appendix I.

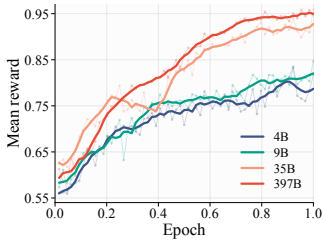


Figure 4: Mean reward during SC-GRPO training at four model scales: 4B, 9B, 35B, and 397B total parameters.

SC-GRPO improves downstream video generation. We finally compare WanPE-397B before and after SC-GRPO using the expert video-evaluation protocol in Section 2.4. Both variants use the same Wan3.0's video generator and differ only in their enhanced conditions. As shown in Table 3, SC-GRPO increases the overall preference score from 42.70 to 49.69 and improves performance across all seven content categories, with larger gains in knowledge, singing and dance, and speech. These results thus show that optimizing semantic consistency also benefits overall downstream video quality, producing more effective textual conditions for video generation.

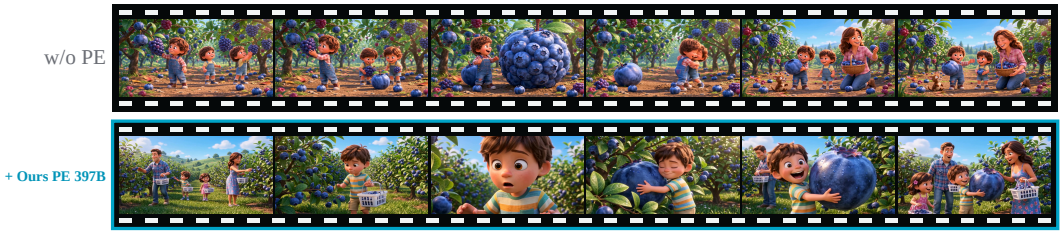


Figure 5: Generated videos with and without PE. The demo is available here.

4 Related Work

Text-to-video generation. Text-to-video generation has progressed rapidly with diffusion and transformer architectures, from early latent video models (Blattmann et al., 2023; Wang et al., 2023; Guo et al., 2024; Ma et al., 2025) to large diffusion transformers such as Open-Sora (Zheng et al., 2024), CogVideoX (Yang et al., 2025), HunyuanVideo 1.5 (Wu et al., 2025), Wan2.2 (Wan Team et al., 2025), LTX series (HaCohen et al., 2026), Lingbot-Video (Ma et al., 2026), and MiniMax-H3 (MiniMax Team, 2026). Closed-source systems include Sora 2 (OpenAI, 2025), Kling (Kling AI, 2026), Seedance 2.5 (ByteDance Seed, 2026), Wan3.0 (Wan Team, 2026).

Prompt enhancement for text-to-video generation. Recent works have explored T2V prompt enhancement from several directions. POS (Ma et al., 2024) jointly optimizes noise and textual prompts with semantic-preserving rewriting, while RAPO (Gao et al., 2025) combines retrieval-augmented modifiers with LLM rewriting to better align user prompts with the training-prompt distribution. VPO (Cheng et al., 2025) and Prompt-A-Video (Ji et al., 2025) train prompt optimizers with supervised and preference-based objectives, while VidPrompter (Wang et al., 2025) adopts multimodal multi-task training with hallucination-aware preference optimization. CAPE-T2V (Jia et al., 2026) studies captioner-anchored two-sided conditioning alignment. Other methods focus on iterative refinement during inference: PhyT2V (Xue et al., 2025) introduces reasoning-based self-refinement for physical realism, SCMAPR (Yang et al., 2026) coordinates scenario-aware agents with semantic verification, and VISTA (Long et al., 2026) iteratively revises prompts using generated-video selection and critique. RAPO++ (Gao et al., 2026) combines training-data alignment, test-time optimization, and rewriter fine-tuning, while PhyPrompt (Wu et al., 2026) trains a physics-aware rewriter with supervised fine-tuning and GRPO. Together, these works highlight the importance of prompt enhancement as an interface between user requests and video generators.

Modern commercial T2V systems also increasingly rely on proprietary prompt enhancers. For example, HappyHorse 1.1 (HappyHorse Team, 2026), Kling 3.0 (Kling AI, 2026), MiniMax-H3 (MiniMax Team, 2026), Seedance 2.0 (Seedance et al., 2026), and Seedance 2.5 (ByteDance Seed, 2026) employ native prompt-enhancement components whose model designs and training recipes remain largely undisclosed. MiniMax-H3 provides H3-Context-IR as a separately accessible prompt enhancer paired with H3-Base, whereas the prompt enhancers of other systems are integrated into their end-to-end generation APIs.

5 Conclusion

This work reframes prompt enhancement for modern video generators. As generators process longer contexts, follow complex instructions, and synthesize videos of up to 30 seconds, the prompt is no longer a decorated caption but the blueprint that orchestrates actions, camera choreography, audio, and narrative across shots and through time before pixel rendering. We introduced WanPE as a scalable realization of this cinematic planning architecture. It learns realized cinematic structure from real-world videos through video-grounded reverse SFT and faithfully preserves user requirements throughout the unfolding plan through SC-GRPO. Scaling WanPE from 4B to 397B consistently improves semantic fidelity and downstream video quality. Ablations show that reverse construction outperforms forward rewriting and that SC-GRPO improves semantic fidelity across model scales, while format-adapted WanPE outputs remain effective across video generators. These findings lay a reliable foundation for the next stage of prompt enhancement development.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, improving readability, formatting tables. All AI-assisted changes were manually reviewed and verified by the authors. Generative AI tools were not used to generate experimental results or make final scientific conclusions. The authors take full responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies prompt enhancement for text-to-video generation. The video corpus used for video-grounded supervision consists of publicly available or licensed videos, and was curated with filtering procedures intended to remove invalid, low-quality, or unsuitable content. Human evaluation was conducted through anonymous pairwise comparisons by expert evaluators, and the collected judgments were used only in aggregate form for research analysis. We recognize that improved cinematic prompt enhancement may amplify both beneficial creative applications and potential misuse of text-to-video systems, including the generation of misleading, harmful, biased, or rights-infringing content. We therefore emphasize that such systems should be deployed with appropriate safety filters, provenance mechanisms, copyright and privacy safeguards, and human oversight. The benchmark and analysis in this paper are intended to improve controllability, semantic fidelity, and evaluation transparency rather than to facilitate harmful generation.

REFERENCES

Artificial Analysis. Video generation benchmarking methodology. URL https://artificialanalysis.ai/video/methodology. Accessed: September 19, 2026.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

ByteDance Seed. Seedance 2.5. https://seed.bytedance.com/en/seedance2_5, 2026.

Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. Vpo: Aligning text-to-video generation models with prompt optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15636–15645, October 2025.

Bingjie Gao, Xinyu Gao, Xiaoxue Wu, Yujie Zhou, Yu Qiao, Li Niu, Xinyuan Chen, and Yaohui Wang. The devil is in the prompts: Retrieval-augmented prompt optimization for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3173–3183, June 2025.

Bingjie Gao, Qianli Ma, Xiaoxue Wu, Shuai Yang, Guanzhou Lan, Haonan Zhao, Jiaxuan Chen, Qingyang Liu, Yu Qiao, Xinyuan Chen, Yaohui Wang, and Li Niu. Rapo++: Cross-stage prompt optimization for text-to-video generation via data alignment and test-time scaling, 2026. URL https://arxiv.org/abs/2510.20206.

Google. Gemini 3.1 pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, 2026.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning, 2024. URL https://arxiv.org/abs/2307.04725.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

HappyHorse Team. Happyhorse 1.1. https://www.happyhorse.cn/, 2026.

Yatai Ji, Jiacheng Zhang, Jie Wu, Shilong Zhang, Shoufa Chen, Chongjian Ge, Peize Sun, Weifeng Chen, Wenqi Shao, Xuefeng Xiao, Weilin Huang, and Ping Luo. Prompt-a-video: Prompt your video diffusion model via preference-aligned llm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18725–18735, October 2025.

Yizhuo Jia, Jingyun Hua, and Yuanxing Zhang. Cape-t2v: Captioner-anchored prompt enhancement toward two-sided conditioning alignment in text-to-video generation. *arXiv preprint arXiv:2608.03046*, 2026.

Kling AI. Kling video 3.0: Professional cinematic ai video production, 2026. URL https://kling.ai/feature/kling-video-3.

Lightricks. Ltx-2.5. https://ltx.io/model/ltx-2-5, 2026.

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

Shijie Ma, Huayi Xu, Mengjian Li, Weidong Geng, Yaxiong Wang, and Meng Wang. Pos: A prompts optimization suite for augmenting text-to-video generation, 2024. URL https://arxiv.org/abs/2311.00949.

Shuailei Ma, Jiaqi Liao, Xinyang Wang, Jingjing Wang, Chaoran Feng, Zijing Hu, Chong Bao, Zichen Xi, Yuqi Gan, Weisen Wang, Yanhong Zeng, Qin Zhao, Zifan Shi, Wei Wu, Hao Ouyang, Qiuyu Wang, Shangzhan Zhang, Jiahao Shao, Yipengjing Sun, Liangxiao Hu, Lunke Pan, Nan Xue, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Ka Leong Cheng. Scaling mixture-of-experts video pretraining for embodied intelligence, 2026. URL https://arxiv.org/abs/2607.07675.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation, 2025. URL https://arxiv.org/abs/2401.03048.

MiniMax Team. Minimax-h3. https://www.minimax.io/blog/minimax-h3, 2026.

OpenAI. Sora 2 is here. https://openai.com/index/sora-2/, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, Jiarui Fang, Jing Fang, Zetao Fang, Chengjian Feng, Yu Gao, Diandian Gu, Dong Guo, Hanzhong Guo, Qiushan Guo, Boyang Hao, Hongxiang Hao, Haoxun He, Jiaao He, Qian He, Tuyen Hoang, Heng Hu, Ruoqing Hu, Yuxiang Hu, Jiancheng Huang, Weilin Huang, Zhaoyang Huang, Zhongyi Huang, Jishuo Jin, Ming Jing, Ashley Kim, Shanshan Lao, Yichong Leng, Bingchuan Li, Gen Li, Haifeng Li, Huixia Li, Jiashi Li, Ming Li, Xiaojie Li, Xingxing Li, Yameng Li, Yiying Li, Yu Li, Yueyan Li, Chao Liang, Han Liang, Jianzhong Liang, Ying Liang, Wang Liao, J. H. Lien, Shanchuan Lin, Xi Lin, Feng Ling, Yue Ling, Fangfang Liu, Jiawei Liu, Jihao Liu, Jingtuo Liu, Shu Liu, Sichao Liu, Wei Liu, Xue Liu, Zuxi Liu, Ruijie Lu, Lecheng Lyu, Jingting Ma, Tianxiang Ma, Xiaonan Nie, Jingzhe Ning, Junjie Pan, Xitong Pan, Ronggui Peng, Xueqiong Qu, Yuxi Ren, Yuchen Shen, Guang Shi, Lei Shi, Yinglong Song, Fan Sun, Li Sun, Renfei Sun, Wenjing Tang, Boyang Tao, Zirui Tao, Dongliang Wang, Feng Wang, Hulin Wang, Ke Wang, Qingyi Wang, Rui Wang, Shuai Wang,

Shulei Wang, Weichen Wang, Xuanda Wang, Yanhui Wang, Yue Wang, Yuping Wang, Yuxuan Wang, Zijie Wang, Ziyu Wang, Guoqiang Wei, Meng Wei, Di Wu, Guohong Wu, Hanjie Wu, Huachao Wu, Jian Wu, Jie Wu, Ruolan Wu, Shaojin Wu, Xiaohu Wu, Xinglong Wu, Yonghui Wu, Ruiqi Xia, Xin Xia, Xuefeng Xiao, Shuang Xu, Bangbang Yang, Jiaqi Yang, Runkai Yang, Tao Yang, Yihang Yang, Zhixian Yang, Ziyan Yang, Fulong Ye, Bingqian Yi, Xing Yin, Yongbin You, Linxiao Yuan, Weihong Zeng, Xuejiao Zeng, Yan Zeng, Siyu Zhai, Zhonghua Zhai, Bowen Zhang, Chenlin Zhang, Heng Zhang, Jun Zhang, Manlin Zhang, Peiyuan Zhang, Shuo Zhang, Xiaohe Zhang, Xiaoying Zhang, Xinyan Zhang, Xinyi Zhang, Yichi Zhang, Zixiang Zhang, Haiyu Zhao, Huating Zhao, Liming Zhao, Yian Zhao, Guangcong Zheng, Jianbin Zheng, Xiaozheng Zheng, Zerong Zheng, Kuan Zhu, and Feilong Zuo. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Wan Team. Wan AI: Leading ai video generation model. https://wan.video, 2026.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Jiapeng Wang, Chengyu Wang, Jun Huang, and Lianwen Jin. Hallucination-aware prompt optimization for text-to-video synthesis. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 10198–10206. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/1133. URL https://doi.org/10.24963/ijcai.2025/1133. AI, Arts & Creativity.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models, 2023. URL https://arxiv.org/abs/2309.15103.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, Linus, Patrol, Peizhen Zhang, Peng Chen, Penghao Zhao, Qi Tian, Songtao Liu, Weijie Kong, Weiyan Wang, Xiao He, Xin Li, Xinchi Deng, Xuefei Zhe, Yang Li, Yanxin Long, Yuanbo Peng, Yue Wu, Yuhong Liu, Zhenyu Wang, Zuozhuo Dai, Bo Peng, Coopers Li, Gu Gong, Guojian Xiao, Jiahe Tian, Jiaxin Lin, Jie Liu, Jihong Zhang, Jiesong Lian, Kaihang Pan, Lei Wang, Lin Niu, Mingtao Chen, Mingyang Chen, Mingzhe Zheng, Miles Yang, Qiangqiang Hu, Qi Yang, Qiuyong Xiao, Runzhou Wu, Ryan Xu, Rui Yuan, Shanshan Sang, Shisheng Huang, Siruis Gong, Shuo Huang, Weiting Guo, Xiang Yuan, Xiaojia Chen, Xiawei Hu, Wenzhi Sun, Xiele Wu, Xianshun Ren, Xiaoyan Yuan, Xiaoyue Mi, Yepeng Zhang, Yifu Sun, Yiting Lu, Yitong Li, You Huang, Yu Tang, Yixuan Li, Yuhang Deng, Yuan Zhou, Zhichao Hu, Zhiguang Liu, Zhihe Yang, Zilin Yang, Zhenzhi Lu, Zixiang Zhou, and Zhao Zhong. Hunyuanvideo 1.5 technical report, 2025. URL https://arxiv.org/abs/2511.18870.

Shang Wu, Chenwei Xu, Zhuofan Xia, Weijian Li, Lie Lu, Pranav Maneriker, Fan Du, Manling Li, and Han Liu. Phyprompt: Rl-based prompt refinement for physically plausible text-to-video generation, 2026. URL https://arxiv.org/abs/2603.03505.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18826–18836, June 2025.

Chengyi Yang, Pengzhen Li, Jiayin Qi, Aimin Zhou, Ji Wu, and Ji Liu. SCMAPR: Self-correcting multi-agent prompt refinement for complex-scenario text-to-video generation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6942–6973, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.345. URL https://aclanthology.org/2026.findings-acl.345/.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025. URL https://arxiv.org/abs/2408.06072.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL https://arxiv.org/abs/2412.20404.

APPENDIX

A VIDEO DATA COLLECTION AND CURATION

This section describes the real-world video corpus for video-grounded supervision. Section A.1 introduces the collection scope. Section A.2 describes the shot- and sequence-level organization. Section A.3 presents the quality filtering and curation criteria. Section A.4 summarizes the corpus.

A.1 COLLECTION SCOPE

Cinematic planning requires supervision that captures both fine-grained realization within individual shots and narrative organization across multiple shots. We therefore adopt a dual-track collection paradigm that retains high-quality shot-level clips and coherent multi-shot sequences. The former provides precise and aesthetically expressive examples of individual cinematic concepts, while the latter captures narrative continuity, editing rhythm, and temporal progression.

The corpus covers ten content dimensions, including narrative, motion, music video, character dialogue, animation, world knowledge, advertising and promotional content, general scenes, procedural demonstrations, and documentaries. These dimensions provide diverse examples of cinematic expression, audiovisual presentation, and narrative structure across production domains.

The source videos are publicly available or licensed for use and span diverse cinematic domains. Long-form videos are divided into clips of no more than 30 seconds, allowing each sample to preserve detailed audiovisual content while remaining suitable for structured cinematic captioning.

A.2 SHOT- AND SEQUENCE-LEVEL ORGANIZATION

At the shot level, the corpus follows a single-shot quality paradigm. Each clip is selected as a high-quality realization of a specific concept under Narrative Core guidance, teaching the model to render a meaningful cinematic shot with narrative and aesthetic coherence. The relevant signals include character actions and emotions, regional action styles, single- and multi-person interactions, lighting and color, camera configuration and movement, music, speech, sound effects, animation-specific appearance and motion, and motion-graphics design. These clips provide dense supervision for semantic precision, visual composition, and aesthetic realization within an individual shot.

At the sequence level, the corpus follows a multi-shot quality paradigm. The emphasis shifts from attributes within shots to relationships among ordered shots. Retained sequences expose changes in action, scene, camera, audio, and visual style, together with the transitions, editing patterns, and narrative relations that connect them. These sequences provide direct supervision for narrative continuity, editing rhythm, and the organization of meaningful shots into a coherent story.

A.3 QUALITY FILTERING AND CURATION

We first filter the collected clips for technical validity, visual quality, and motion quality. The remaining samples are then curated according to their temporal structure and cinematic value.

For shot-level clips, curation emphasizes semantic clarity, aesthetic quality, and the completeness of the central event. The criteria cover character performance, including narrative-core actions, regional action styles, and emotional expressions across single- and multi-person interactions. They also cover visual language, including mood-linked lighting, color, composition, camera configuration, and camera movement. Additional criteria address audiovisual performance, animation-specific styles and motion, and motion-graphics design for diverse cinematic production contexts.

For sequence-level clips, curation emphasizes narrative continuity and editing rhythm. A retained sequence should be understandable as an independent unit, and its ordered shots should form a meaningful narrative progression. We further consider the richness of its cinematic organization, including changes in action, scene, camera, audio, transitions, editing structure, and pacing. Content-specific criteria are also applied where appropriate, including cross-shot visual-style consistency for animation and effective information delivery in motion-graphics sequences.

A.4 Resulting Corpus

The resulting corpus contains approximately $1.05\mathrm{M}$ clips of up to 30 seconds across ten content dimensions. By combining fine-grained shot-level evidence with realized multi-shot organization, the corpus supports learning cinematic conditions at both local and long-range temporal scales. Appendix B describes how these videos are converted into video-grounded cinematic targets.

B Video-Grounded Reverse Construction

This section describes the reverse-construction procedure used to generate cinematic caption targets. We first specify the hierarchical structure shared by all targets in Section B.1. We then describe the category-specific captioning requirements in Section B.2 and the distinct annotation focuses for single-shot and multi-shot data in Section B.3. Finally, Section B.4 summarizes the caption-quality checks applied before the targets are retained for supervised fine-tuning of the prompt enhancer.

B.1 Hierarchical Caption Structure

Each cinematic target y_i follows a shared hierarchy with a video-level description and shot-level descriptions. The video-level description summarizes the overall content, setting, visual style, narrative perspective, pacing, and audio design. The shot-level descriptions decompose the video into temporally ordered shots with explicit timestamps. Each shot specifies its composition, subjects, actions, lighting, camera movement, transitions, dialogue, music, and sound effects. The description also records how these elements progress over time within the temporally ordered video.

B.2 Category-Specific Captioning Requirements

The category-specific instruction s_i adjusts the descriptive emphasis within the shared caption structure. Examples include general scenes, large-motion content, music videos, rich-text content, animation, advertisements and motion graphics, visual effects, and complex camera movements.

The *general* category covers, for example, everyday life, natural scenery, and social activities. Its captions emphasize comprehensive scene coverage, accurate subject descriptions, and the completeness of relevant visual and audiovisual details across diverse real-world environments and contexts.

The *large-motion* category covers sports, racing, martial arts, acrobatics, and dance. Its captions describe displacement, velocity, body-posture dynamics, action continuity, and fast-paced audio semantics to support complex motion understanding. The *complex-camera-movement and long-take* category includes push, pull, pan, tilt, tracking, and continuous shots. Its captions systematically describe camera movement type, direction, speed, spatial trajectory, and narrative function.

The *music-video and music-BGM* category covers musical performances, dance routines, and music videos with close audiovisual coupling. Its captions jointly describe performer expressiveness, movement rhythm, music style, audiovisual synchronization, music IP, and lyrical content.

The *rich-text* category covers subtitles, signs, handwriting, screen recordings, posters, user interfaces, and other videos containing dense or structurally complex text. Its captions transcribe on-screen text and describe its position, style, layout, temporal changes, and relationship to visual subjects. The *animation* category covers stylized content such as 2D hand-drawn, cel-shaded, and 3D-rendered animation. Its captions identify fine-grained visual styles, animation-specific motion primitives, and audio characteristics for style-conditioned video generation across animation forms.

The *advertisement and motion-graphics* category covers commercial films, brand videos, promotional content, and program packaging. Its captions focus on visual design language, motion pacing, information delivery, and brand presentation. The *visual-effects* category covers post-production content such as particle effects and digital compositing. Its captions describe the effect type, visual appearance, temporal evolution, and integration with real or rendered scenes.

B.3 Single-Shot and Multi-Shot Caption Focuses

For single-shot data, the annotation focuses on semantic precision and aesthetic quality within an individual shot. Character performance includes narrative-core actions, regional action styles, and emotional expressions in single-person and multi-person interactions. Visual language includes mood-linked lighting and color as well as basic and advanced camera patterns. Audio-visual performance includes music, character-centric vocal archetypes, and diegetic or stylized sound effects. The annotation also covers multi-dimensional 2D and 3D animation styles, animation-specific motion primitives, and structured character voice profiles. Motion graphics and packaging follow a three-axis taxonomy of application scenario, motion effects, and visual style.

For multi-shot data, the annotation focus shifts from within-shot attributes to inter-shot relationships and temporal dynamics. The captions describe narrative continuity and editing rhythm across the ordered shots. This single-shot to multi-shot progression follows the distinction between describing a narratively meaningful shot and describing how multiple shots form a coherent story.

B.4 Caption Quality Control

We assess the generated captions for structural completeness, timestamp validity, and consistency with the source video. We regenerate captions that fail these checks. The retained captions serve as the video-grounded targets used for supervised fine-tuning of the prompt enhancer.

C Semantic-Consistency Reward and SC-GRPO

C.1 Semantic-Consistency Reward

Given a user request x and an enhanced cinematic condition y, the semantic-consistency reward evaluates whether y preserves the requirements specified by x and remains consistent across shots and over time. Additional details are accepted when they remain compatible with the request.

We use Qwen3.7-Max as a text-only evaluator to produce $r_{\mathrm{sem}}(x, y) \in [0, 100]$, where a higher score indicates more complete and accurate requirement preservation. The evaluation covers nine dimensions:

- **Style:** visual medium, genre, realism, aesthetic treatment, mood, and requested presentation style.
- **Subjects:** subject identity, count, appearance, attributes, roles, and persistence across shots.
- **Actions:** requested actions, state changes, interactions, event order, and action–subject bindings.
- **Dialogue:** spoken content, wording constraints, language, speaker identity, and speaker–utterance bindings.
- **Sound:** music, vocals, ambience, sound effects, and their requested timing or relation to visual events.
- **Camera:** shot scale, viewpoint, movement, composition, transitions, shot count, and camera sequencing.
- **Lighting:** illumination, color temperature, contrast, direction, source, and temporal changes in lighting.
- **Spatial relations:** relative positions, orientations, containment, contact, and subject–object relations.
- **Scene:** setting, time, weather, environment, props, and continuity of scene-level constraints.

C.2 Scoring Criteria and Evaluator Output

The evaluator treats the user request as the sole source of required constraints. Semantically equivalent paraphrases are accepted, as are elaborations compatible with the explicit constraints and se-

mantic relations entailed by the request. Added details are penalized only when they omit, weaken, replace, or contradict a requested element or disrupt consistency across the sequence.

Discrepancies are assessed according to their semantic severity and scope. The evaluator checks for omitted requirements, altered subjects or actions, unsupported substitutions, incorrect subject–attribute or speaker–dialogue bindings, temporal and shot-order errors, and contradictions within or across shots. A localized mismatch involving a secondary detail receives a smaller penalty than a violation of the central subject, action, narrative progression, or an explicit production constraint. The final score reflects a holistic judgment across the nine dimensions.

A score of 100 denotes complete semantic preservation across all specified requirements. Scores in $(60, 100)$ denote usable conditions with one or more identifiable discrepancies, whereas scores in $[0, 60]$ denote failures involving substantial omission, alteration, or conflict.

For each request–condition pair, we report three fields: `score`, an integer in $[0, 100]$; `primary issue`, the dimension containing the most consequential discrepancy or `None`; and `rationale`, a concise, evidence-grounded explanation of the judgment. The scalar score is used as $r_{\text{sem}}(x, y)$, while the diagnostic fields support inspection and error analysis.

Human verification. We manually inspect 200 prompt-enhancer outputs and their corresponding Qwen3.7-Max reward judgments. Human assessment agrees with the evaluator on 194 cases, yielding an agreement rate of 97% and supporting its reliability for automated evaluation.

C.3 Semantic-Consistency Reward Demonstrations

The following cases show how the reward treats complete preservation, a localized production-constraint violation, and a conflict with a central requested action. Each case includes the user request and enhanced condition. We report the score, issue, and rationale from Qwen3.7-Max. In the penalized cases, underlined text marks the output spans responsible for the deduction.

C.3.1 Perfect Case

User request (10 seconds).

> A dark magician loses control while manipulating dark magic. Use multiple shots to show the pain of the backlash from the failure. In the end, he finds a way to end the farce.

Enhanced condition.

> The video consists of 4 shots. This dark fantasy film scene depicts a dark magician losing control while manipulating dark magic, suffering a painful backlash, and ultimately ending the farce. The scene is set in a dark, cavernous chamber with high, arched stone ceilings and rough-hewn walls. The overall visual style features a gritty, high-contrast texture that emphasizes the raw power of the magic. The narrative is presented from a third-person objective perspective, utilizing a cinematic shooting style with a rapid editing pace to build tension. The audio landscape is composed of a tense, ominous orchestral score and a rich layer of sound effects, including magical crackles and pained vocalizations.
>
> Shot 1 [00.00-02.00s]. A medium shot at eye level frames the scene with a shallow depth of field, focusing on the dark magician. The scene is set at night, with the lighting dominated by hard, cool-toned magical light, presenting an extremely dark low-key characteristic with strong light-dark contrast. The color saturation is low, overlaid with a strong cool-toned grading, jointly creating a dark and tense emotional atmosphere. The key light source is the hard, cool-toned magical light emanating from his hands, forming an underlight effect on the dark magician, presenting a high lighting ratio. In the center of the frame, the dark magician, a man in his late 30s to early 40s with a pale complexion, sharp cheekbones, a prominent nose, and a short, dark beard and mustache, stands facing forward. He has long, straight, jet-black hair that falls past his shoulders and wears a floor-length, deep charcoal gray robe with a high collar and subtle, dark embroidery on

the sleeves. He is positioned in the midground, with the background consisting of the dark, cavernous chamber featuring high, arched stone ceilings and rough-hewn walls, where faint, flickering torchlight is visible in the distance. He is performing a spell. Standing with his arms outstretched and palms facing upward, he manipulates a swirling mass of dark, smoky energy between his hands, which emits a deep purple glow. His brow is furrowed in concentration, and his eyes are fixed on the energy. The camera slowly dollies in, tightening the frame on him as the dark energy intensifies, accompanied by a low, humming drone and the sound of crackling, ethereal energy, while a tense, ominous orchestral score featuring low strings and a deep brass undertone begins to swell.

Shot 2 [02.00-04.50s]. The intensifying energy drives a hard cut to a close-up of the magician's face, captured at eye level with a very shallow depth of field. The key light source remains the hard, cool-toned magical light, now flaring erratically, forming an underlight effect on the dark magician, presenting a high lighting ratio. The dark magician's face fills the frame, positioned slightly off-center to the left, with the background completely blurred. The dark energy erupts violently. The purple light flares erratically, casting harsh, moving shadows across his face. His expression shifts from concentration to shock and pain; his eyes widen, and his mouth opens in a silent gasp. The camera, with a slight handheld shake, pans slowly to the right, following his head as he recoils slightly. The sound design escalates with a loud, sharp crackle and a deep, explosive boom, while the orchestral score surges in volume and intensity, adding dissonant strings and powerful percussion hits.

Shot 3 [04.50-07.00s]. The violent recoil triggers a hard cut to a low-angle medium shot, emphasizing the magician's loss of control. The key light source is the hard, cool-toned magical light, which continues to flare, forming an underlight effect on the dark magician, presenting a high lighting ratio. The dark magician is positioned in the center of the frame, with the background showing the dark, cavernous chamber. He is thrown backward by the force of the magical backlash. His body arches back, his head tilted up, and his arms flail uncontrollably. His face is contorted in agony, with his eyes squeezed shut and his mouth open in a pained grimace. The camera, with a noticeable handheld shake, tilts down slowly, following his body as he stumbles backward. The sound of a sustained, high-pitched energy whine mixes with a low, rumbling bass tone, and the magician lets out a sharp, pained gasp, followed by a low, guttural groan. The orchestral score reaches a chaotic crescendo with frantic string passages and heavy, pounding drums.

Shot 4 [07.00-10.00s]. A hard cut jumps to a static, eye-level close-up, capturing the resolution of the event. The key light source is the hard, cool-toned magical light, which quickly dissipates, leaving the scene in near darkness, forming an underlight effect on the dark magician, presenting a high lighting ratio. The dark magician's face is centered in the frame, with the background completely blurred. He ends the spell. His expression shifts from pain to grim determination. With a sudden, decisive movement, he clenches his right fist, and the chaotic energy instantly vanishes with a sharp, final "whoosh" sound. The lighting returns to the dim, ambient torchlight of the chamber. He then lowers his head, his eyes closing briefly before opening with a look of weary resolve. He takes a sharp intake of breath, and the orchestral score abruptly cuts out, replaced by a single, low, resonant note that fades into silence. Finally, he looks directly forward, his expression somber, having ended the farce.

Qwen3.7-Max judgment.

Score: 100. **Primary issue:** None. **Rationale:** Preserves the subject, action narrative, and shot requirements. Added scene, lighting, sound, and camera details remain compatible. No semantic omission, weakening, or conflict is identified.

C.3.2 ACCEPTABLE CASE

User request (15 seconds).

> On a vast open field, two armies of soldiers are arrayed in formation, charging at each other. War drums thunder as the two armies instantly clash and become a tangled mass of fighting. The brutality and solemn tragedy of the war are emphasized through multiple cuts between different shots and camera distances, accompanied by corresponding battle sound effects. After the battle, bodies are scattered everywhere, with severed limbs and remains strewn across the sandy battlefield. On the screen, the text appears: 'A single general's success is built on ten thousand bleaching bones.' The visuals are required to be realistic, and the shot transitions should be smooth.

Enhanced condition.

> The video consists of 5 shots. This film scene, an ancient epic large-scale battlefield battle clip, depicts two armies clashing on a vast field, resulting in a brutal battle and a tragic aftermath. The scene is set on a vast, flat, sandy plain under a pale, overcast sky. The overall visual style features a gritty, realistic texture. The narrative is presented from an objective third-person perspective, utilizing a cinematic shooting style with a pace that accelerates into the battle before slowing for the conclusion. The audio landscape is composed of a somber, epic orchestral score and a detailed layer of sound design, including war drums, battle cries, and the sounds of combat.
>
> Shot 1 [00.00-02.00s]. A high-angle long shot establishes the scene, its deep focus capturing the entire expanse of the sandy beige field under a pale, overcast sky. The lighting is dominated by soft, diffused daylight, creating a somber atmosphere. The frame is organized in a symmetrical composition, with two massive armies of soldiers arrayed in dense, rectangular formations on the left and right sides of the frame, facing each other across a wide central gap. The soldiers are clad in dark, earth-toned armor and helmets, carrying spears and shields. In the background, a low, hazy horizon line separates the field from the sky. The camera holds a steady, static position. The two armies begin to charge. From their static positions, the soldiers on both sides simultaneously break into a full run, surging forward toward the center of the frame. As they charge, a powerful, rhythmic drumming of war drums begins, accompanied by a low, ominous orchestral string drone that builds in volume, establishing a tense and foreboding atmosphere.
>
> Shot 2 [02.00-03.50s]. The forward momentum of the charge drives a hard cut to an eye-level medium shot, plunging the viewer into the action. The camera, with a shallow depth of field, focuses on the front ranks of the charging soldiers, who fill the frame, running directly toward the camera. The soldiers, wearing dark, earth-toned armor and helmets, carry spears and shields. Their faces are contorted in fierce, aggressive shouts, with mouths wide open, eyes glaring intensely, and brows deeply furrowed, projecting a unified, explosive anger as they charge. The camera slowly dollies in, moving forward with the advancing army, enhancing the sense of overwhelming force. The thunderous drumming intensifies, now joined by the chaotic roar of thousands of battle cries and the thudding sound of thousands of feet running on sand, while the orchestral score swells with driving percussion and brass, amplifying the ferocity of the charge.
>
> Shot 3 [03.50-06.00s]. The collision of the armies triggers a hard cut to an eye-level medium shot at the heart of the conflict. The frame is a chaotic, claustrophobic mass of fighting soldiers, with a shallow depth of field isolating small clusters of combatants. The soldiers, still in their dark, earth-toned armor, are locked in close-quarters combat, wielding swords and shields. They engage in a brutal melee, their bodies a tangled mass of violent motion. The camera, exhibiting a slight handheld shake, pans slowly to the right, sweeping across the carnage. The audio erupts into a cacophony of clashing metal, guttural grunts, and sharp, per-

cussive impacts of weapons striking shields and armor, while the orchestral score reaches a frantic, dissonant crescendo, mirroring the chaos of the battle.

Shot 4 [06.00-08.00s]. A hard cut jumps from the melee to a stark, high-angle extreme long shot, revealing the devastating scale of the battle. The frame is filled with a sprawling, chaotic pattern of fighting soldiers scattered across the sandy field, with no clear lines or formations. The camera slowly dollies out, pulling back to reveal even more of the battlefield, emphasizing the vastness of the conflict and the sheer number of combatants. The chaotic sounds of battle continue, but become slightly more distant, matching the wider perspective, while the intense, dramatic orchestral music continues, sustaining the high-stakes atmosphere.

Shot 5 [08.00-15.00s]. The dolly-out of the previous shot transitions via a hard cut to a grim, high-angle long shot of the aftermath. The camera holds a steady, static position, looking down upon the battlefield, which is now littered with the bodies of fallen soldiers. The sandy ground is strewn with severed limbs and remains, and the surviving soldiers are seen standing amidst the carnage. The chaotic sounds of battle abruptly cease, replaced by a low, mournful wind and the faint, distant groans of the wounded. The orchestral score shifts to a slow, somber, and melancholic theme played by strings, underscoring the tragedy. As the camera holds on this desolate scene, white text in a serif font fades in at the center of the screen, reading: 'A single general's success is built on ten thousand bleaching bones.' The text remains on screen as the shot concludes, providing a final, solemn commentary on the cost of war.

Qwen3.7-Max judgment.

Score: 80. **Primary issue:** Camera. **Rationale:** Repeatedly replaces smooth transitions with hard cuts.

C.3.3 FAILURE CASE

User request (10 seconds).

Shoot a narrative MV with a healing, slice-of-life feel for a deep and soothing Chinese male-vocal folk ballad. Design four scenes with smooth transitions, presenting different vignettes of life in the streets and alleys late at night. Use fade-in and fade-out transitions to connect the figures of rushing passersby. The editing pace should match the deep, soothing melody, the emotional tone should be gentle and healing, and the shot transitions must be precisely synchronized with the musical beats.

Enhanced condition.

The video consists of 4 shots. This narrative music video, with a healing, slice-of-life genre, presents a series of vignettes showing people in a modern Chinese city at night. The setting is a series of interconnected urban streets and alleys at night. The overall visual style features a soft, slightly grainy texture that enhances the cinematic feel. The narrative perspective is third-person objective, captured with a smooth, stabilized camera, and the editing pace is slow and contemplative, matching the music. The audio features a deep, soothing Chinese male-vocal folk ballad, characterized by a gentle, mature male voice singing with a clear, warm timbre, accompanied by a sparse acoustic guitar melody, creating a gentle, healing, and introspective atmosphere.

Shot 1 [00.00-02.50s]. The video opens with a medium shot, framed at eye level, capturing a wet city street at night. The scene is illuminated by the warm, golden glow of streetlights, which cast long reflections on the dark, wet asphalt. The composition uses the leading lines of the road to draw the eye toward a young woman in her early 20s, of East Asian descent with a fair skin tone, who is positioned in the center of the frame, facing away from the camera. She has long, straight, black hair and wears a light beige trench coat over a white t-shirt, dark blue jeans, and white sneakers. In the background, blurry pedestrians

walk along the sidewalk, and a red traffic light is visible in the distance. The focus is shallow, keeping her sharp while the background is softly blurred. The young woman walks slowly forward down the center of the street. Initially seen from behind, she walks at a steady, unhurried pace, her arms swinging gently by her sides, her head held straight as she looks down the street. As she walks, the camera slowly dollies in, smoothly tracking her movement and gradually tightening the frame around her. A gentle, melodic acoustic guitar intro begins, followed by the entrance of a deep, soothing male vocal, singing a gentle, flowing melody in a clear, warm, and mature tone. The music establishes a calm, introspective mood that mirrors the quiet, solitary walk.

Shot 2 [02.50-05.00s]. A fade-out and fade-in transition, synchronized with a musical beat, shifts the scene from the open street to a narrow, dimly lit alley. The shot is a medium shot, framed at eye level, with a shallow depth of field. The alley is flanked by brick walls and features a single, warm-toned lantern hanging from a building on the right, casting a soft, golden light. In the midground, a middle-aged man in his 40s, of East Asian descent with a medium skin tone, stands on the left side of the frame, facing right in profile. He has short, black hair and wears a dark gray jacket over a light-colored shirt. He holds one small, round, white food item, likely a steamed bun, in his hands. The man eats the steamed bun. Holding the bun with both hands, he brings it to his mouth and takes a bite, chewing slowly with a relaxed expression. As he eats, the camera slowly trucks left, subtly shifting the perspective. The male vocalist continues to sing, his voice maintaining its gentle, soothing quality, with the lyrics flowing seamlessly across the transition, linking the two scenes emotionally.

Shot 3 [05.00-07.50s]. Another fade-out and fade-in transition, timed to a musical beat, moves the view to a different part of the alley. This medium shot, framed at eye level, focuses on an elderly woman in her 70s, of East Asian descent with a light skin tone, sitting on a low step on the left side of the frame, facing right in a 3/4 profile. She has short, gray hair and wears a dark, patterned blouse. She is knitting with one pair of thin, metal knitting needles and one ball of light gray yarn. The scene is lit by a warm, soft light from an unseen source, highlighting her hands and face. The elderly woman knits. With a focused, serene expression, she deftly manipulates the knitting needles, her hands moving in a rhythmic, practiced motion as she works on a piece of light gray fabric. The camera slowly dollies in, moving closer to her and emphasizing the intricate, repetitive action. The male vocal continues, the melody remaining consistent and soothing, underscoring the quiet, meditative quality of her activity.

Shot 4 [07.50-10.00s]. A final fade-out and fade-in transition, synchronized with a musical beat, returns to a wider view of the alley. This medium shot, framed at eye level, captures a group of three people walking away from the camera down the center of the alley. The group consists of a man in his 30s, of East Asian descent with a medium skin tone, wearing a dark jacket and dark pants, a woman in her 30s, of East Asian descent with a fair skin tone, wearing a light-colored coat, and a child in their 5s, of East Asian descent with a fair skin tone, wearing a bright yellow jacket. The alley is dimly lit by distant streetlights, creating a soft, hazy atmosphere. The group walks away together. The man and woman walk side-by-side, with the child between them, holding the woman's hand. They move at a leisurely pace, their figures receding into the distance. The camera slowly dollies out, pulling back and widening the frame as they walk away. The male vocal continues, the song maintaining its gentle, melancholic tone, providing a sense of closure as the figures disappear into the night.

Qwen3.7-Max judgment.

Score: 45. **Primary issue:** Action. **Rationale:** Replaces rushing passersby with slow, leisurely, and relaxed figures, creating major conflicts in both subject attributes and the core action.

Table 4: Training configurations for WanPE.

Supervised Fine-Tuning		**SC-GRPO Post-Training**	
Configuration	Value	Configuration	Value
Training epochs	2	Training epochs	1
Peak learning rate	2×10^{-5}	Learning rate	2×10^{-6}
Weight decay	0.1	Rollouts per request	12
Maximum sequence length	23,040	KL penalty coefficient	0.01
		Temperature	1.0

D Implementation and Training Details

This section summarizes the training configuration of WanPE, which consists of video-grounded supervised fine-tuning followed by SC-GRPO post-training. The left and right parts of Table 4 report the hyperparameters for supervised fine-tuning and SC-GRPO, respectively.

E WanPEval Design and Construction

This section presents the distribution of WanPEval and representative requests at three granularity levels. We first describe the coverage of the 249 requests in Section E.1. We then present intent-level, scene-level, and shot-level examples across multiple content categories and target durations in Section E.2. Finally, Section E.3 describes the interface used for blind pairwise video evaluation.

E.1 Benchmark Coverage and Distribution

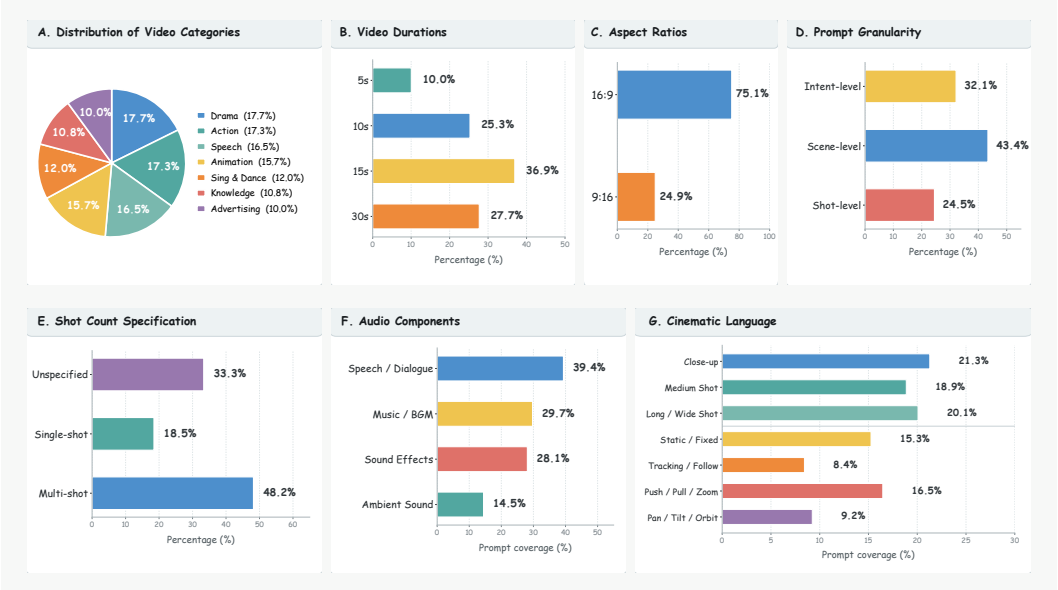


Figure 6: Distribution of the 249 requests in WanPEval. The panels report category, duration, aspect ratio, prompt granularity, shot-count specification, audio components, and cinematic language.

Figure 6 summarizes the distribution of the 249 requests in WanPEval. The benchmark covers seven content categories, four target durations from 5 to 30 seconds, two aspect ratios, and three request-granularity levels. It further characterizes shot-count specifications, audio components, and cinematic language across the requests. Together, these dimensions reflect practical text-to-video generation scenarios and requirements for comprehensive and controlled model evaluation.

E.2 Representative Requests Across Granularity Levels

The nine requests are taken from WanPEval, with three examples at each granularity level. They illustrate the diversity of the benchmark across categories, durations, and levels of request detail.

Intent-level examples.

Drama, 10 seconds, 16:9.

> Cinematic realistic style, a company corridor in the early morning, a mid-long shot of an intern working overtime alone, constantly complaining, being overheard by the boss.

Speech, 5 seconds, 16:9.

> Generate a historical court drama dialogue video with 3 shots. The content is about an imperial concubine aggressively scolding a maid, and the maid responding aggrievedly.

Animation, 15 seconds, 9:16.

> 3D exquisite realistic animation, in front of a Gothic church under the moonlight, a little girl in a Gothic nun's attire holds a thorny rose and makes eye contact with a plaster statue of Jesus, cool color tone, exquisite and beautiful.

Scene-level examples.

Drama, 10 seconds, 16:9.

> A young man with the power to sense and purify evil energies walks alone through a deserted alley on a rainy night. Suddenly, he is entangled by a ferocious evil spirit lurking in the dark. He faces the roaring, pouncing evil spirit alone, fighting against the malevolent forces by himself to protect the peace of the human world.

Action, 30 seconds, 16:9.

> A wandering monk wearing a bamboo hat is traveling when he witnesses a young woman being harassed by thugs. He immediately rushes forward, uses the bamboo flute in his hand to fight them off and rescues the woman. After accepting her thanks, he turns and walks away, playing the bamboo flute as he goes.

Animation, 15 seconds, 16:9.

> Classic fairytale princess style 3D animation. In a glass vase in a flower shop, a small white flower quietly blooms, and a glowing flower fairy appears from within the petals. She slowly floats among the flower stands, looking around at the room full of fresh flowers and water droplets. Suddenly, she hears the shop's doorbell ring, immediately returns inside the flower bud, and the flower bud closes again.

Shot-level examples.

Drama, 10 seconds, 16:9.

> Shot 1: Long shot, modern realistic style, a playground. Afternoon sun casts dappled light and shadows through tree leaves, with students walking and jogging in the background. The camera slowly pushes in, focus gradually shifts from the light spots on the playground floor to the running students in the distance. The foreground leaves are slightly blurred to create a sense of spatial depth.
>
> Shot 2: Mid shot, tracking. The male protagonist, carrying a backpack, walks towards the school building. The camera follows from his side-rear, with a stabilizer for smooth tracking. A slight high angle captures the rhythm of his steps and his elongated shadow. The background is blurred to make the subject stand out.

> Shot 3: Mid shot pushing in to a close-up. The female protagonist is sitting on a bench, reading a book. Dappled sunlight falls on the tips of her hair and the book pages. The camera slowly pushes in from a mid shot to a side-profile close-up. A slight handheld shake is added for a more lifelike feel. A shallow depth of field focuses on the female protagonist's expression, with the background softly blurred.
>
> Shot 4: Close shot. The male protagonist walks up to the female protagonist and hands her a bottle of water. The camera pans horizontally from the male protagonist's hand to the female protagonist's hand, keeping both centered in the frame. A slight focus pull transitions between them, capturing the moment their hands touch. The background remains slightly blurred.

Speech, 5 seconds, 16:9.

> A dialogue scene of a long-awaited reunion in an old mansion from the Republic of China era. The text and tone must exactly match the instructions. Shot 1: A two-person mid-shot, static camera. A girl in a simple and elegant cheongsam (in the calm tone of a young lady from a wealthy family): "You still know to come back." A man in a dark Zhongshan suit (in the sincere tone of a mature and steady young man): "I haven't forgotten our promise." Shot 2: Cut to a close-up of the girl's side profile. The girl (emotionally overwhelmed, crying, voice trembling and choked): "All these years, I've been waiting for you."

Speech, 30 seconds, 16:9.

> On a late autumn evening, in a quiet street alley, a tall ginkgo tree gently sheds its golden leaves. Pale, warm evening light filters through the branches, casting mottled bright spots. The ambient sound is of a low wind and the subtle rustling of leaves scraping the ground. The camera uses a fixed medium shot, where a woman in a simple, light-colored knitted sweater with her long hair hanging naturally stands quietly by the ginkgo tree, her gaze fixed on the end of the alley. Then, the frame slowly cuts to a close-up of her fingers gently twirling a leaf, capturing the small and restrained action. Next, it slowly transitions to a wide shot; the woman turns slightly to the side and walks with a composed pace deeper into the street alley, her silhouette gradually receding. The camera slowly pulls back, finally stopping on a quiet picture of her back intertwined with the long alley in late autumn. Narrator (in a calm tone): "The wind brushes the long street, yellow leaves fall silently. Those fragments buried deep in the years, when recalled, leave only a touch of warmth and a bit of melancholy that was never spoken."

These examples illustrate how request detail varies across the three granularity levels. Intent-level requests specify a high-level scenario with limited temporal detail. Scene-level requests describe the progression of events without shot-by-shot instructions. Shot-level requests organize content into shots and specify elements such as shot scale, camera movement, focus, action, and dialogue.

E.3 Blind Pairwise Evaluation Interface

Figure 7 shows the blind pairwise evaluation interface. It presents the request above two anonymized candidate videos labeled A and B. The interface provides joint and individual playback controls, a note field for each video, and four overall outcomes: A better, B better, both good, and both poor.

F Evaluation Protocol

This section describes the text-level semantic evaluation, video-generation settings, and blind pairwise expert evaluation used on WanPEval.

F.1 Text-Level Semantic Evaluation

To avoid self-evaluation bias, we do not use the Qwen3.7-Max reward model for evaluation. Instead, Gemini 3.1 Pro and GPT-5.4 independently evaluate the semantic consistency between each user re-

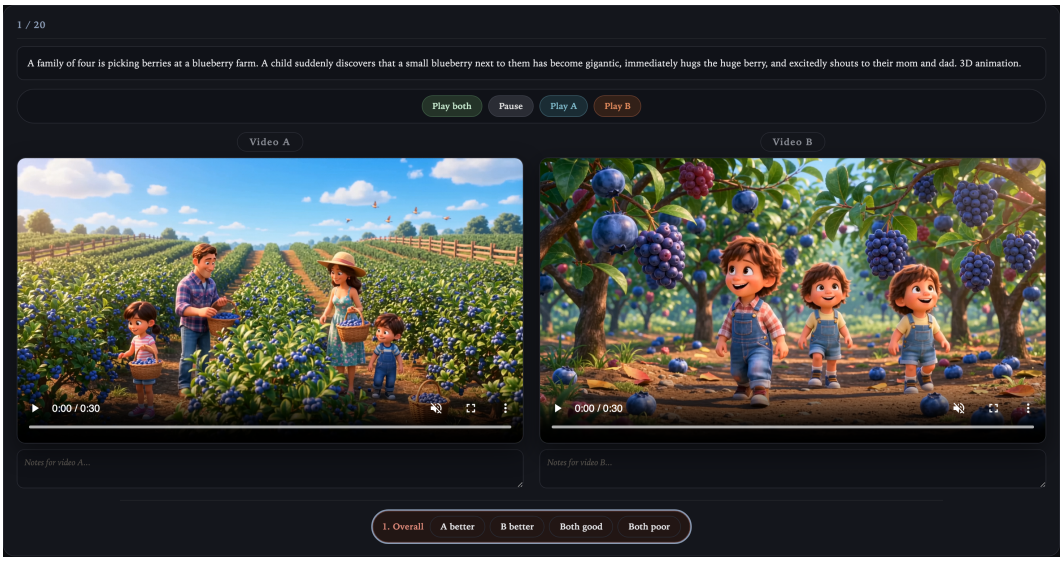


Figure 7: Blind pairwise evaluation interface. The anonymized interface displays the request and two candidate videos with playback controls, note fields, and the four overall evaluation outcomes.

quest x and enhanced condition y using the same nine-dimensional rubric defined in Section C.1. Neither evaluator is involved in SC-GRPO training. Each evaluator returns a score in $[0, 100]$, together with a verdict, primary issue, and dimension-level evidence. We report Gemini 3.1 Pro results in the main paper and GPT-5.4 results as an independent robustness check in the appendix.

An output is classified as perfect only when it receives a score of 100, while a score of 60 or below indicates failure. Section C.3 provides representative perfect, acceptable, and failure cases illustrating the same semantic-consistency criteria.

F.2 Video Generation and Baseline Configuration

Each method generates one video for every applicable request. We pass WanPE conditions to Wan3.0 unless a different setting is specified. The original-request baseline uses Wan3.0 without prompt enhancement. LTX-2.5 and MiniMax-H3 use their native prompt enhancers before their base generators. End-to-end API baselines receive the original request through their default native prompt-processing pipelines. These settings preserve the intended configuration of each external baseline. They also use the same generator for WanPE and the original-request baseline.

Videos are generated at 720p or 768p according to each baseline's supported resolution. Target duration and aspect ratio follow the corresponding WanPEval request whenever supported. We fix the random seed when seed control is available to improve reproducibility across repeated comparisons.

F.3 Blind Pairwise Expert Evaluation

We conduct anonymous pairwise video comparisons with 60 experts in screenwriting, directing, cinematography, and related film disciplines. The evaluation jointly considers request adherence, perceptual quality, and cinematic coherence.

Evaluator training. Before formal evaluation, evaluators review standardized instructions explaining the evaluation objectives, the four possible outcomes, and the operation of the annotation interface. They then complete a trial evaluation using representative comparison pairs to become familiar with the procedure and clarify how request fulfillment, visual quality, temporal coherence, camera language, and audiovisual presentation should inform the overall judgment.

Evaluation procedure. For each request, all available method pairs are constructed. Method identities are hidden, and the left–right order is randomized. Evaluators view the request and two candi-

Table 5: Baseline-specific preference scores across seven content categories and overall.

Method	Action	Anim.	Speech	Ad.	Sing & Dance	Drama	Know.	Overall
LTX-2.5 vs. Ours								
LTX-2.5	6.78	10.83	36.21	19.23	5.56	16.10	26.32	17.34
Downstream generator: Wan3.0's video generator								
Original Request	43.33	48.33	35.34	38.46	47.37	35.00	11.54	37.85
+WanPE-397B	**50.85**	**57.50**	**40.52**	**48.68**	**52.63**	**51.69**	**50.00**	**50.28**
Kling 3.0 vs. Ours								
Kling 3.0	9.32	10.00	35.34	16.67	26.25	18.97	17.57	18.95
Downstream generator: Wan3.0's video generator								
Original Request	40.00	42.50	26.32	32.05	35.00	32.50	12.82	32.54
+WanPE-397B	**50.00**	**62.50**	**44.92**	**57.89**	**48.75**	**50.86**	**56.58**	**52.84**
HappyHorse 1.1 vs. Ours								
HappyHorse 1.1	12.07	17.50	26.27	31.58	8.75	24.56	35.53	21.71
Downstream generator: Wan3.0's video generator								
Original Request	36.44	40.00	34.75	24.36	47.50	24.14	3.85	31.07
+WanPE-397B	**44.92**	**60.83**	**50.00**	**50.00**	**51.25**	**52.54**	**46.15**	**51.14**
MiniMax-H3 vs. Ours								
MiniMax-H3	11.54	25.00	22.50	33.33	25.00	39.17	42.11	27.83
Downstream generator: Wan3.0's video generator								
Original Request	35.96	37.07	44.07	28.95	37.50	30.00	12.82	33.33
+WanPE-397B	**42.73**	**59.48**	**53.39**	**46.15**	**45.00**	**45.83**	**48.72**	**49.14**
Seedance 2.0 vs. Ours								
Seedance 2.0	**41.11**	44.90	35.96	33.78	29.03	43.75	30.56	37.94
Downstream generator: Wan3.0's video generator								
Original Request	27.36	22.22	38.60	31.58	31.43	22.41	5.26	25.98
+WanPE-397B	38.46	**52.73**	**42.24**	**47.30**	**38.89**	**50.86**	**39.47**	**44.76**

date videos labeled A and B using the interface shown in Figure 7. The interface provides joint and individual playback controls and an optional note field for each candidate.

Evaluators select one of four outcomes: A better, B better, both good, or both poor. They assess each pair holistically based on correspondence to the request and overall video quality.

Aggregation. We aggregate the judgments using the preference and Bradley–Terry scores defined in Section 2.4. The four-outcome protocol distinguishes decisive preferences from ties and differentiates between "both good" and "both poor" judgments.

G Additional Quantitative Results

This section provides additional quantitative results that supplement Table 1 and Figure 3. We report baseline-specific preference results and exact breakdowns by duration and request granularity on the 5-15-second subset of WanPEval.

G.1 Baseline-Specific Preference Results

Table 5 reports preference results relative to external baselines. Each block compares an external system with the original-request baseline and WanPE-397B, with the gray row indicating the downstream generator used by the latter two. Scores aggregate pairwise judgments within each block.

G.2 Fine-Grained Preference Breakdowns

Tables 6 and 7 list the exact values represented by the markers and intervals in the left panel of Figure 3. Each entry follows the order $S_{\text{lower}}/S/S_{\text{upper}}$. WanPE-397B attains the highest central score at every duration and request granularity.

Table 6: Exact preference intervals by duration. Entries are $S_{\text{lower}}/S/S_{\text{upper}}$.

Method	5 seconds	10 seconds	15 seconds
LTX-2.5	17.29 / 18.69 / 20.09	16.04 / 18.47 / 20.90	14.39 / 16.10 / 17.80
Kling 3.0	26.27 / 31.11 / 35.94	19.68 / 23.47 / 27.26	14.07 / 16.10 / 18.12
HappyHorse 1.1	15.67 / 20.28 / 24.88	20.65 / 24.00 / 27.36	23.20 / 26.80 / 30.41
MiniMax-H3	25.00 / 29.09 / 33.17	31.77 / 36.46 / 41.16	33.93 / 38.30 / 42.67
Seedance 2.0	42.11 / 46.20 / 50.29	36.23 / 41.40 / 46.58	39.94 / 44.31 / 48.69
Downstream generator: Wan3.0's video generator			
Original request	39.63 / 46.08 / 52.53	27.70 / 33.63 / 39.57	26.09 / 30.59 / 35.08
+WanPE-4B	41.74 / 48.17 / 54.59	35.83 / 41.27 / 46.70	38.61 / 44.00 / 49.38
+WanPE-9B	43.84 / 50.91 / 57.99	39.32 / 44.40 / 49.47	41.12 / 45.78 / 50.43
+WanPE-35B	42.59 / 50.69 / 58.80	41.84 / 48.76 / 55.67	40.97 / 46.45 / 51.93
+WanPE-397B	48.37 / 56.74 / 65.12	42.37 / 49.91 / 57.45	43.18 / 49.43 / 55.68

Table 7: Exact preference intervals by request granularity. Entries are $S_{\text{lower}}/S/S_{\text{upper}}$.

Method	Intent-level	Scene-level	Shot-level
LTX-2.5	17.23 / 19.21 / 21.19	12.92 / 14.39 / 15.87	15.70 / 18.18 / 20.66
Kling 3.0	17.35 / 20.66 / 23.98	16.45 / 18.79 / 21.12	20.44 / 24.03 / 27.62
HappyHorse 1.1	21.18 / 25.42 / 29.65	21.78 / 25.00 / 28.22	20.51 / 23.74 / 26.97
MiniMax-H3	33.28 / 38.37 / 43.45	31.05 / 35.05 / 39.05	30.87 / 34.84 / 38.80
Seedance 2.0	41.61 / 46.83 / 52.05	37.89 / 42.21 / 46.53	35.10 / 39.23 / 43.36
Downstream generator: Wan3.0's video generator			
Original request	29.07 / 33.89 / 38.70	27.86 / 32.29 / 36.72	28.53 / 35.87 / 43.21
+WanPE-4B	39.97 / 45.96 / 51.95	37.84 / 42.05 / 46.25	34.95 / 41.67 / 48.39
+WanPE-9B	42.90 / 48.28 / 53.66	40.29 / 44.87 / 49.45	38.01 / 43.53 / 49.06
+WanPE-35B	41.94 / 49.33 / 56.72	42.83 / 48.35 / 53.86	38.75 / 44.44 / 50.14
+WanPE-397B	44.82 / 52.44 / 60.06	43.91 / 49.26 / 54.61	40.98 / 49.32 / 57.65

H Cross-Generator Transfer and Format Adaptation

Table 2 quantitatively evaluates cross-generator transfer. This section complements those results with a qualitative trace of the format-adaptation step. Starting from the same canonical output of WanPE-397B, GPT-5.4 serializes the cinematic plan into the explicit multimodal fields used by H3-Context-IR and the compact prose format used by LTX-2.5-PE. The following case allows the planning content to be inspected before and after these representation changes.

H.1 Retro Anime Mecha at Sunset

User request (Animation, 15 seconds, intent-level).

> 80s old anime feel, a giant mecha and a young girl standing on a seaside breakwater, the sunset is especially red, with film grain, classic Japanese sci-fi anime flavor, very nostalgic, like an old OVA animation accidentally seen on late-night TV.

Original WanPE-397B output.

> The video consists of 4 shots. This film scene is a 2D Cartoon belonging to the Retro-Style Anime genre, with a style evocative of classic 80s Japanese sci-fi anime OVAs. The art style features a Semi-Realistic overall design, characterized by Clean Line Art that defines the forms, while space is constructed using Conventional Linear Perspective and volume is rendered through Cel Shading(1-3 Tones). The story depicts a young girl and a giant mecha standing together on

a seaside breakwater at sunset. The scene is set on a long, concrete breakwater extending into the sea, with the weathered, blocky structure's edge facing the open ocean. A massive, humanoid mecha stands on the left side of the breakwater, its body angled slightly towards the sea, and a young girl stands to its right, near the center of the frame. The background is a vast, open ocean under a fiery red-orange sunset sky, with the sun low on the horizon. The video scene is set at sunset, the lighting overall led by soft Sunset Warm Light, the image's light-dark contrast is medium-low, presenting a dim mid-tone characteristic. Color saturation is high, overlaid with a strong intensity Warm Deep Orange filter, the whole leaning toward a warm tone tendency, together creating a nostalgic and melancholic emotional atmosphere. The visual texture has a distinct film grain, mimicking an old OVA animation. The narrative is presented from a third-person objective perspective, utilizing a static camera and a slow, measured editing pace. The audio landscape combines ambient environmental sounds, a key sound effect, and a nostalgic synthesizer score to create a melancholic and nostalgic atmosphere.

Shot 1 [00.00-03.50s]. A static, eye-level long shot with a deep depth of field establishes the scene, featuring a prominent film grain effect. The composition is balanced, with the massive, dark gray, blocky humanoid mecha on the left and a small, 16-year-old girl in a simple white dress on the right, both silhouetted against a fiery red-orange sunset. The two subjects stand motionless on the concrete breakwater. The mecha, with its weathered armor showing signs of wear, is positioned on the left side of the frame, its form a dark silhouette against the vibrant sky. To its right, the girl stands near the center of the frame, her hair gently blowing in the wind. The setting sun casts a brilliant, fiery red-orange glow across the entire scene, reflecting off the calm ocean surface. The audio is composed of the gentle, rhythmic lapping of waves against the breakwater and the distant, low cry of a single seagull, as a nostalgic, melancholic 80s-style synthesizer melody begins to play softly in the background.

Shot 2 [03.50-07.00s]. A hard cut transitions to a static, low-angle medium shot, focusing on the girl. The camera is positioned below her, looking up, which emphasizes the immense scale of the mecha looming over her. The deep depth of field keeps both the girl and the mecha's lower torso in focus. The girl, with her natural, no-makeup appearance, is in the foreground, her back partially to the camera, her face turned towards the sea. Her expression is pensive and calm, her gaze fixed on the horizon. The mecha's massive, dark gray leg and torso dominate the left side of the frame, its sheer size dwarfing her. The sunset continues to cast a warm, red-orange light, catching the edges of her white dress and the mecha's weathered armor. The ambient sound of lapping waves continues, and the synthesizer melody swells slightly in volume, enhancing the nostalgic mood.

Shot 3 [07.00-11.00s]. Following a hard cut, the scene shifts to a static, low-angle close-up on the mecha's head. The frame is filled with the mecha's helmet, its dark gray metal showing visible scratches and worn paint. The deep depth of field allows the details of the weathered armor to be clear. The intense red-orange light of the sunset reflects brightly off the mecha's large, dark visor. The mecha performs a subtle action. Initially, its head is still. Then, it slowly turns its head, a faint, low-frequency mechanical whirring sound accompanying the movement. Its dark visor reflects the intense red-orange light of the sunset. As this happens, the synthesizer melody holds a sustained, resonant note, creating a moment of poignant stillness.

Shot 4 [11.00-15.00s]. A final hard cut transitions to a static, eye-level extreme long shot, pulling back to show the entire scene once more. The composition is a wide, panoramic view, with the breakwater, the mecha, and the girl all appearing small against the vast, fiery red-orange sky and the darkening sea. The deep depth of field captures the entire vista. The mecha and the girl are now small figures on the breakwater, their silhouettes stark against the dramatic sunset. The camera remains static, holding on this final, expansive view. The ambient sound of lapping waves continues, and the synthesizer melody slowly fades out, leaving only the

sound of the waves and the lingering image of the two figures against the dying light.

Format-adapted output for MiniMax-H3.

integrated_multimodal_description: [Shot 1] 2D-animated, Retro-Style Anime, long shot, static, eye-level angle, establishing a long concrete breakwater extending into the open sea at sunset. The visual style is evocative of classic 80s Japanese sci-fi anime OVAs, with a semi-realistic overall design, clean line art, conventional linear perspective, cel shading in 1–3 tones, and distinct film grain. A massive, dark gray, blocky humanoid mecha with weathered armor showing signs of wear stands on the left side of the breakwater, its body angled slightly toward the sea. To its right near the center of the frame, a 16-year-old girl in a simple white dress stands motionless beside it, her hair gently blowing in the wind. Both are silhouetted against a vast open ocean and a fiery red-orange sunset sky with the sun low on the horizon. Soft sunset warm light creates medium-low contrast and dim mid-tones, while high saturation and a strong warm deep orange filter produce a nostalgic, melancholic atmosphere. [Shot 2] At 0:03.500 cut to 2D-animated, Retro-Style Anime, medium shot, static, low-angle framing focused on the girl. The deep depth of field keeps both the girl in the foreground and the mecha's lower torso in focus. The girl, shown with a natural no-makeup appearance and her back partially to camera, turns her face toward the sea with a pensive, calm expression, her gaze fixed on the horizon. The mecha's massive dark gray leg and torso dominate the left side of the frame, looming over her and emphasizing its immense scale. The warm red-orange sunset light catches the edges of her white dress and the mecha's weathered armor. [Shot 3] At 0:07.000 cut to 2D-animated, Retro-Style Anime, close-up, static, low-angle framing on the mecha's head. The frame is filled by the mecha's helmet, its dark gray metal rendered with visible scratches and worn paint. Deep depth of field keeps the weathered armor details clear. The intense red-orange sunset reflects brightly across the mecha's large dark visor. Initially the head is still, then it slowly turns, the visor continuing to mirror the fiery light. [Shot 4] At 0:11.000 cut to 2D-animated, Retro-Style Anime, extreme long shot, static, eye-level angle, pulling back to a wide panoramic view of the entire breakwater against the vast sea and sky. The mecha and the girl appear as small silhouettes on the long, weathered concrete structure, stark against the expansive fiery red-orange sunset and the darkening ocean. Deep depth of field captures the full vista. The camera remains static, holding the final melancholic composition as the light fades toward dusk.

overall_soundscape: Gentle, rhythmic waves lap against the concrete breakwater throughout, with soft water movement echoing lightly against the weathered blocky structure. A distant, low cry of a single seagull punctuates the opening atmosphere over the open ocean air. In the second shot, the seaside ambience remains sparse and steady, with the breeze implied by the girl's softly moving hair and dress. In the close-up on the mecha, a faint, low-frequency mechanical whirring accompanies the slow turn of its head, subtle and restrained against the otherwise quiet shoreline. In the final wide shot, the waves continue as the dominant diegetic layer, leaving the scene with only sea ambience and open coastal space.

non_diegetic_music: A nostalgic, melancholic 80s-style synthesizer melody begins softly under the opening shot. The melody swells slightly in Shot 2, then holds a sustained, resonant note during the mecha's head turn in Shot 3. In the final shot, the synthesizer slowly fades out, leaving only the sound of the waves.

Format-adapted output for LTX-2.5.

An eye-level long shot opens in the style of a classic 1980s Japanese sci-fi anime OVA, rendered as a semi-realistic 2D cartoon with clean line art, conventional linear perspective, cel shading in a few bold tones, and visible film grain, framing a long concrete breakwater stretching into the sea at sunset where a massive,

> weathered dark gray humanoid mecha stands on the left and a small sixteen-year-old girl in a simple white dress stands to its right near the center, both motionless and silhouetted against a fiery red-orange sky reflected across the calm ocean, while gentle waves lap against the structure, a distant seagull cries, and a nostalgic melancholic synthesizer melody begins softly; a hard cut shifts to a static low-angle medium shot of the girl in the foreground with her back partly to camera and her face turned toward the horizon, her hair moving lightly in the sea breeze as the mecha's immense leg and lower torso loom above her on the left, the warm deep orange sunset light catching the edges of her dress and the worn armor plating, emphasizing the scale difference and her calm, pensive stillness as the synth swells slightly; another hard cut moves into a static low-angle close-up of the mecha's head, its helmet filling the frame with scratched paint, weathered metal, and a large dark visor reflecting the intense sunset glow, and after a moment of stillness the head slowly turns with a faint mechanical whir, the sustained synth note deepening the poignant atmosphere; a final hard cut resolves into an eye-level extreme long shot that pulls the entire scene back into a panoramic silhouette, the breakwater, girl, and mecha reduced to small figures against the vast open sea and blazing red-orange sky, the image holding steady in nostalgic melancholy as the synthesizer fades and only the rhythmic wash of waves remains.

In this example, format adaptation changes the representation rather than the shot organization. The four-shot progression from the establishing view to the girl, the mecha's head turn, and the final panoramic composition remains intact. H3 separates visual content, diegetic sound, and music into explicit fields, whereas LTX-2.5 compresses the same progression into continuous prose.

I From SFT to SC-GRPO

The largest downstream gains from SC-GRPO occur in knowledge, singing and dance, and speech. These categories share cross-modal and temporal dependencies. Speech requires dialogue to remain aligned with speakers, actions, expressions, camera presentation, and timing. Singing and dance requires coordination among lyrics, music, choreography, camera movement, and editing rhythm. Knowledge requires explanations to follow the order of visual demonstrations and on-screen content. Such dependencies create opportunities for binding drift in long cinematic conditions. By jointly evaluating semantic consistency across nine dimensions, SC-GRPO reinforces these relationships throughout the shot sequence, leading to particularly strong gains in these three categories.

I.1 Case Studies

The following case studies illustrate these cross-modal and temporal dependencies. In each case, SFT preserves the broad request but alters constraints assigned to particular shots or stages. SC-GRPO preserves the same cinematic structure while maintaining those long-range bindings.

I.1.1 Dawn-to-Night Alley Narration

User request.

> Generate a narrated video of an alley from dawn to night. The background music is soothing piano music throughout, accompanied by faint ambient sounds of pedestrian footsteps, vendors' soft cries, etc. Narration tone: a gentle and clear female voice, with a light and slow pace, transitioning with the scene from dawn to night. The text, punctuation, and pauses must be followed exactly, fitting the warm and everyday tone.
>
> Slow push-in camera movement. Morning fog lightly envelops the small town's streets and alleys, with gentle morning light spreading across the land. Gradually focus on the alley entrance, transitioning to a close-up shot. The bluestone path is moist with morning fog, and the low walls on both sides are clearly outlined. With narration: "The morning light first appears, the mortal world begins to wake."
>
> [Scene 2] Cut to a medium shot with a horizontal pan, capturing the shops on both sides and pedestrians hurrying through.

[Scene 3] Cut to a close-up shot, capturing the busy hand movements of a grilled cold noodle vendor, accompanied by the sizzling sound of the noodles. With narration: ”The smoke and fire of human life, an ordinary day.”

[Scene 4] Cut to a long shot with a static camera, capturing the sunset dyeing the sky, with layers of evening clouds reflected on the river surface.

[Scene 5] Fade to black transition. A close-up shot shows the alley as night falls and the decorative lights turn on. The alley is bustling with voices.

[Scene 6] Cut to a long shot. The deep night is quiet and vast. The alley has shed its daytime clamor, with sparse figures. The evening breeze gently sweeps through the entire alley. The scene is quiet and soft. With narration: ”The night is peaceful, and the heart finds its rest.”

WanPE-397B-SFT output.

The video consists of 6 shots. This serene daily life timelapse short film of a historic ancient town alley presents a visual and auditory journey through a modern Chinese water town, observing its transition from dawn to night as the lives of its inhabitants unfold within its alleys. The entire scene is set in an exterior narrow alleyway paved with uneven gray stone slabs. The narrative perspective is third-person observational, employing smooth drone or crane shots and stabilized camera work, edited at a slow, contemplative pace. The audio landscape features a gentle female narrator speaking English with a calm tone, accompanied by ambient environmental sounds and a continuous, soothing piano score that creates a peaceful atmosphere.

Shot 1 [00.00-02.75s] From a high angle, an extreme long shot establishes the scene at dawn, where soft cool-toned blue-hour light forms a front flat light effect on the misty street, presenting weak light-dark contrast, mid-tone characteristics, low color saturation, and a strong analogous colors grading. A thick layer of pale sky-blue fog blankets the town, creating a balanced composition with the central alley acting as a leading line into the hazy distance. In this static frame, the foreground shows the tops of buildings with dark charcoal-gray tiled roofs flanking the alley, while two indistinct human silhouettes are positioned in the center midground. Two distant figures slowly walk forward along the mist-shrouded old stone street in the early dawn haze, gradually moving away from the camera towards the far end of the street. As they move, a gentle, mature-sounding female narrator’s voice begins, stating, ”The morning light first appears,” delivered at a leisurely pace with a warm, clear tone over a newly introduced, simple, and gentle piano melody that sets a tranquil mood.

Shot 2 [02.75-04.92s] A hard cut brings the viewer down to an eye-level medium full shot inside the alley, now filled with bright, diffuse daylight filtering through the thinning fog, where volumetric sunny light forms a back light effect on pedestrians walking in the alley, presenting a high lighting ratio. The focus is shallow, sharpening on the immediate cobblestone path while blurring the background. On the left side of the frame, a building facade displays one rectangular wood signboard painted deep red with gold characters reading ”Xiuxiang Pavilion” above a window, and one vertical fabric banner colored white with black calligraphy reading ”Zhou Family...” hanging next to it. Several blurry pedestrians are scattered throughout the midground. Multiple pedestrians walk briskly across the sunlit old stone street; some head towards the left side of the frame, others move towards the right side of the frame, all traveling quickly past traditional storefronts on both sides of the road. Simultaneously, the camera executes a rapid dolly-in combined with a slow pan to the left, sweeping the view past the architecture. As this movement occurs, the narrator completes the thought from the previous shot: ”the mortal world begins to wake.” The gentle piano music continues unchanged, providing a steady emotional throughline.

Shot 3 [04.92-07.62s] Following the implied motion, a hard cut lands on an eye-level close-up, focusing tightly on a moment of creation under soft ordinary daylight that forms a top light effect on the chef’s hands, presenting a medium lighting ratio. The depth of field is extremely shallow, isolating the action against a completely blurred background. In the center of the frame, a pair of tanned-skinned hands belonging to a food vendor wearing a crisp white apron works over one large, circular, flat-topped metal griddle colored matte black. One small bowl-shaped woven bamboo basket colored natural tan sits out of focus in the foreground. The food vendor uses one wide, flat metal spatula featuring a silver blade and a brown wooden handle to swiftly spread one mound of cooked noodles mixed with chopped green vegetables, colored off-white and vibrant spring-green, evenly across the hot iron pan surface, then lifts the spatula after finishing the spreading motion. This action is accompanied by a sharp, distinct sizzling sound. Over this image of culinary craft, the

narrator observes, "The smoke and fire of human life, an ordinary day," connecting the mundane act to the larger rhythm of existence as the piano music plays on.

Shot 4 [07.62-09.88s] A stark hard cut transitions the scene to sunset, revealing a static, eye-level long shot of a river. Soft sunset warm light forms a silhouette effect on the bridge and buildings, presenting a high lighting ratio, while providing a rim light. The composition is perfectly symmetrical, with a multi-arched stone bridge centered in the upper half of the frame, its form reflected flawlessly in the still water below. The sky behind it burns with the fiery orange and muted lavender hues of dusk. Small silhouetted figures can be seen standing on the bridge in the background. Several silhouetted pedestrians stroll slowly across the arched stone bridge spanning the calm river at sunset, their dark outlines clearly visible against the glowing orange-purple twilight sky. The uninterrupted piano melody bridges the temporal shift, suggesting a continuity of peace despite the dramatic change in scenery.

Shot 5 [09.88-12.08s] Cutting from the open expanse of the river, the scene shifts to an eye-level close-up looking out from darkness into the same alley at night. Hard LED light forms a front light effect on the hanging lanterns, presenting a high lighting ratio. The shot is framed by a dark foreground element, creating a window onto the midground where multiple cylindrical paper lanterns with wood/metal frames, colored scarlet-red and golden-yellow, hang illuminated on the walls. These lights cast a warm glow and create beautiful bokeh circles in the shallow depth of field. Blurry figures populate the background. Crowds of people wander casually along the lantern-lit old stone street at night, moving slowly among glowing decorative installations hung on both sides of the alley. The camera performs a very slight, slow pan to the left, subtly shifting the arrangement of lights. The constant piano theme accompanies this intimate, festive view, maintaining the video's serene core.

Shot 6 [12.08-15.00s] A final hard cut pulls back to a high-angle long shot, concluding the journey. The alley is now bathed in the cool, deep blue light of late evening, punctuated by the warm yellow-orange glow of wall-mounted lamps and the multiple cylindrical paper lanterns lining the path. Volumetric incandescent light forms a complex light direction effect on the street environment, presenting a high lighting ratio, while mixed background light provides a rim light. The composition mirrors the opening shot but is grounded in the tangible reality of the lit alley rather than abstract fog. A few solitary figures are present in the midground and background. Sparse pedestrians walk slowly along the dimly lit quiet old stone street late at night, most heading towards the far end of the alley, leaving only a few lingering figures near the foreground. As the camera begins a slow, steady dolly-out, pulling away from the scene, the narrator delivers the final line: "The night is peaceful, and the heart finds its rest." The piano music swells slightly with the narration before gently fading out, bringing the tranquil observation to a close.

WanPE-397B with SC-GRPO.

The video consists of 6 shots. This narrated video, a slice-of-life piece set in a modern Chinese town, documents the passage of a day from dawn to night in a narrow alley. The scene is set in a narrow, straight pedestrian alley. The overall visual style features a soft, slightly hazy texture. The narrative perspective is third-person objective, captured with a stable, cinematic shooting style and a slow, contemplative editing pace. The audio landscape is composed of a gentle, clear female voiceover, a continuous soothing piano score, and ambient sounds that shift from quiet morning stillness to bustling daytime activity and finally to nighttime quiet.

Shot 1 [00.00-03.50s] The video opens with an eye-level long shot, transitioning to a close-up, that slowly pushes in on a foggy alley at dawn. The scene is bathed in the cool, diffused light of early morning, with a soft, warm glow breaking through the mist. The composition is centered, with the alley's bluestone path creating strong leading lines that draw the eye into the frame. The alley is flanked by two-story buildings with off-white plaster walls and dark gray tiled roofs, featuring traditional Chinese architectural elements like upturned eaves and wooden window frames. The ground is paved with irregular, dark gray bluestone slabs, appearing damp and reflective. On the left side of the alley, there are low stone walls and planters with green foliage, while the right side features shopfronts. The background shows the alley receding into the distance, partially obscured by fog. A gentle, soothing piano melody begins, accompanied by the faint, distant sound of a single footstep on stone. The camera slowly pushes in, the focus shifting from the wide, fog-shrouded view to a close-up of the alley entrance, where the damp, dark gray bluestone path and the low, off-white walls are clearly defined. A gentle, clear female voiceover, with a light and slow pace, speaks: "The morning light first appears, the mortal world begins to wake."

> Shot 2 [03.50-05.00s] A hard cut transitions to an eye-level medium shot. The camera pans horizontally to the right, capturing the now-bustling alley in the soft, even light of mid-morning. The frame is filled with the facades of shops on both sides, their signs and awnings visible, and pedestrians walking briskly in both directions, their forms slightly blurred by motion. The soothing piano music continues, now layered with the ambient sounds of multiple footsteps on stone and the indistinct, soft cries of vendors.
>
> Shot 3 [05.00-07.50s] A hard cut jumps to an eye-level close-up, focusing on the hands of a grilled cold noodle vendor. The shot is static, with a shallow depth of field that isolates the action. The vendor's hands, moving with practiced speed, use metal tongs and a spatula to expertly manipulate noodles on a hot, flat griddle. The sizzling sound of the cooking noodles is prominent, and steam rises from the griddle. The gentle, clear female voiceover returns, with a light and slow pace, speaking: "The smoke and fire of human life, an ordinary day." The soothing piano music continues underneath.
>
> Shot 4 [07.50-09.50s] A hard cut shifts to a static, eye-level long shot of the sunset. The sky is a gradient of warm orange and deep purple, with layers of evening clouds reflected on the surface of a river in the foreground. The scene is tranquil, with the soft, fading light of dusk. The soothing piano music continues, its tone becoming more serene and reflective.
>
> Shot 5 [09.50-11.50s] A fade to black transitions to an eye-level close-up of the alley as night falls. The shot is static, with a shallow depth of field. The alley is now illuminated by warm, decorative lights strung along the buildings, casting a golden glow on the dark gray bluestone path. The alley is bustling with people, their voices creating a lively ambient soundscape. The soothing piano music continues, its melody soft and warm.
>
> Shot 6 [11.50-15.00s] A hard cut pulls back to a static, eye-level long shot of the alley at deep night. The scene is quiet and vast, with the alley now mostly empty, save for a few sparse figures. The warm decorative lights continue to illuminate the path, but the overall atmosphere is one of quietude. The gentle, clear female voiceover, with a light and slow pace, speaks: "The night is peaceful, and the heart finds its rest." The soothing piano music continues, its melody soft and resolving, as the sound of a gentle evening breeze sweeps through the alley.

Gemini 3.1 Pro scores the SFT and SC-GRPO outputs at 50 and 100, respectively. The SFT output changes the opening push-in and close-up into a static extreme long shot, omits the fade-to-black transition, removes specified vendor calls and nighttime crowd sounds, and weakens the specified food-vendor action. After SC-GRPO, the six-stage morning-to-night progression, three narration segments, camera transitions, and evolving ambient sound remain bound to their intended scenes.

I.1.2 PHOTOSYNTHESIS EXPLANATION

User request.

> Generate a video explaining the key concepts of photosynthesis, presented in a single continuous shot (one-shot). The screen should simultaneously display conceptual text and the overall reaction formula. Use intuitive scenes of plant leaves and chloroplasts to correspond with the formula. The video should be pure visuals, with no real people on screen. The visual content must be strongly synchronized with the narration, and the style should be simple and rigorous.

WanPE-397B-SFT output.

> The video consists of 1 shot. This 3D educational animated short film about the photosynthesis process, set in a modern, abstract digital space, visually explains the process of photosynthesis by showing a plant leaf transforming sunlight, water, and carbon dioxide into glucose and oxygen. The narrative is presented from an objective perspective, utilizing smooth CGI animation to create a slow, deliberate pace. The audio landscape features a gentle ambient piano score and clear voiceover narration that guides the viewer through the scientific process. Shot 1 [00.00-15.00s] The sequence begins with a high-angle extreme long shot, featuring deep depth of field and a balanced composition under soft volumetric sunny light that creates a front-side light effect on the green leaves with a medium lighting ratio, while background daylight provides a rim light. In this vast dark void, two large, curved, vibrant emerald green plant leaves hang suspended; one occupies the upper-left quadrant, arcing downwards towards the center, and the second,

smaller leaf is positioned below it in the lower-right quadrant. A thin, horizontal beam of bright white light stretches across the midground between them. As a calm narrator's voice with a velvety texture begins, stating, "Photosynthesis is the process," accompanied by a gentle, sparse piano melody establishing a serene mood, the camera starts a slow, steady dolly-in movement toward the larger upper leaf. Simultaneously, translucent pale blue gaseous vapor rises slowly from the bottom edge of the frame, and faint, shimmering white particles drift down from the top. Small, glowing pure white text reading "Sunlight" fades into view above the light beam, as the narrator continues, "by which plants use sunlight." The camera's inward journey continues, transitioning the frame to a close-up, then pushing further until it becomes an eye-level extreme close-up focused on the leaf's surface, where hexagonal cells containing small, oval-shaped lime green chloroplasts are revealed. The narrator explains, "Chlorophyll in chloroplasts captures light energy," as the camera seamlessly passes through the cell membrane into the cytoplasm, finally settling on a single chloroplast in the center of the frame. Inside, stacked disc-like thylakoid structures glow with a warm golden-yellow light, and tiny white spheres representing protons flow along a chain of proteins, illustrating the conversion of light energy into chemical energy as the narrator says, "converting it into chemical energy." Following this, the camera rapidly pulls back out of the cell, returning to an eye-level full shot of the entire leaf against the black void. Here, the complete chemical equation for photosynthesis materializes in glowing pure white text across the center of the screen: "6CO + 6HO $\rightarrow$ CHO + 6O", with the word "Sunlight" written above the reaction arrow. The narrator concludes, "This energy is used to convert carbon dioxide and water into glucose and oxygen." To illustrate this final stage, multiple transparent spherical bubbles containing smaller, self-illuminated white spheres float up from the leaf's surface, while other transparent bubbles containing hexagonal molecular structures also rise, visually representing the production of oxygen and glucose before drifting away into the darkness as the shot ends.

WanPE-397B with SC-GRPO.

The video consists of 1 shot. This educational explanatory video explaining the key concepts of photosynthesis presents a visual and auditory explanation of the process, where a camera moves through a plant leaf to reveal the chemical reaction within its cells. The scene is set in a natural, outdoor environment featuring a plant leaf. The overall visual style features a clean, sharp, and high-contrast photographic quality. The narrative perspective is objective, utilizing a smooth, digitally animated camera movement within a single, continuous shot to create a deliberate, educational pace. The audio features a clear, calm voiceover narration with a steady, medium speech rate and a gentle, informative tone, accompanied by a subtle, ambient electronic background music that enhances the scientific atmosphere. Shot 1 [00.00-15.00s]. The shot opens with a medium shot at an eye-level angle, with a deep depth of field, presenting a balanced composition of a vibrant green plant leaf positioned in the center of the frame against a softly blurred background of other foliage. The scene is illuminated by bright, natural daytime sunlight, creating a warm, clear atmosphere. The leaf, with its detailed venation, is the central focus. The camera then begins a slow, smooth, and continuous dolly-in movement, pushing forward into the leaf's surface. As the camera advances, the narrator's calm, clear voice begins, stating, "Photosynthesis is the process by which plants convert light energy into chemical energy." Simultaneously, the text "PHOTOSYNTHESIS" appears in a clean, white, sans-serif font at the top-center of the screen, and the chemical formula "6CO + 6HO + Light Energy $\rightarrow$ CHO + 6O" fades in below it in a matching white, sans-serif font. The camera's movement transitions the view from the macroscopic leaf to a microscopic level, revealing a cross-section of the leaf's internal structure, with the camera moving through the spongy mesophyll layer. The narrator continues, "This process occurs in the chloroplasts," as the camera zooms into a single plant cell, focusing on a bright green, oval-shaped chloroplast. The narrator explains,

Table 8: Additional semantic-consistency results evaluated by GPT-5.4. † denotes results computed on the 5–15-second subset.

Semantic consistency evaluated by GPT-5.4							
Method	Overall↑	5 seconds↑	10 seconds↑	15 seconds↑	30 seconds↑	Perfect↑	Failure↓
Fwd. Rewriting	61.0	62.6	63.1	60.3	59.6	0.8	57.4
LTX-2.5-PE	70.0	72.4	74.4	66.0	70.3	2.8	39.8
H3-Context-IR	64.3†	66.0	64.6	63.6	–	5.6†	57.5†
Ours							
WanPE-4B-SFT	32.9	43.0	31.4	33.2	30.2	0.0	93.6
+ SC-GRPO	64.2 +31.3	73.2 +30.2	62.2 +30.8	62.4 +29.2	65.2 +35.0	2.0 +2.0	53.0 −40.6
WanPE-9B-SFT	35.2	40.9	30.2	40.5	30.8	0.0	92.8
+ SC-GRPO	61.3 +26.1	70.6 +29.7	62.7 +32.5	60.7 +20.2	57.5 +26.7	2.0 +2.0	55.8 −37.0
WanPE-35B-SFT	42.3	49.0	40.1	42.6	41.4	0.0	85.1
+ SC-GRPO	74.6 +32.3	79.0 +30.0	72.1 +32.0	**76.9** +34.3	72.1 +30.7	5.6 +5.6	33.3 −51.8
WanPE-397B-SFT	46.1	49.4	43.3	50.8	41.4	0.0	81.9
+ SC-GRPO	**78.3** +32.2	**84.3** +34.9	**79.0** +35.7	76.3 +25.5	**78.4** +37.0	**7.2** +7.2	**26.1** −55.8

> "Inside the chloroplasts, chlorophyll captures sunlight," and the camera pushes further into the chloroplast, revealing its internal thylakoid membranes, which are shown absorbing golden-yellow light particles. The narrator adds, "and uses this energy to split water molecules, releasing oxygen as a byproduct," while the camera shows water molecules (HO) entering the chloroplast and oxygen molecules (O) being released. Finally, the camera pulls back slightly to show carbon dioxide (CO) entering the chloroplast and glucose (CHO) being synthesized, as the narrator concludes, "Carbon dioxide is then fixed into glucose, a sugar that provides energy for the plant." Throughout the shot, a subtle, ambient electronic background music plays, enhancing the scientific atmosphere, and the visual elements remain perfectly synchronized with the narration, providing a clear, rigorous, and intuitive explanation of photosynthesis.

Gemini 3.1 Pro scores the SFT and SC-GRPO outputs at 60 and 100, respectively. The SFT output replaces the default photographic presentation with a fully animated digital setting. After SC-GRPO, the one-shot structure, conceptual text, reaction formula, plant-to-chloroplast progression, and synchronization with narration remain consistent with the request.

I.2 Additional Results with GPT-5.4

We additionally evaluate semantic consistency using GPT-5.4. As shown in Table 8, SC-GRPO improves the overall score by 26.1–32.3 points across all model scales, with consistent gains at every duration. It also increases the perfect rate and substantially reduces the failure rate. The 397B model achieves the highest overall score of 78.3 and the lowest failure rate of 26.1, further confirming the effectiveness of SC-GRPO under an independent evaluator.

J Qualitative Comparison with Existing Prompt Enhancement Methods

This section presents textual conditions generated by different methods for the same user requests. We compare the open-source methods RAPO Gao et al. (2025), VPO Cheng et al. (2025), SCMaPR Yang et al. (2026), and Prompt-A-Video Ji et al. (2025) instantiated with OpenSora Zheng et al. (2024) and CogVideoX Yang et al. (2025) against WanPE-4B across varied subjects, durations, and levels of request granularity.

J.1 INTENT-LEVEL REQUESTS

J.1.1 TAI CHI SWORD IN AN ANTIQUE COURTYARD

User request (Action, 5 seconds, intent-level).

> A one-shot take of a modern Chinese traditional style enthusiast practicing Tai Chi sword alone in an antique courtyard during the morning light. The movements are slow and continuous, including actions like thrusting and cleaving with the Tai Chi sword, focusing on the aesthetic beauty of the martial art itself and the practitioner's concentration. The background music is a soft guzheng melody, accompanied by the sound of the sword blade cutting through the air and light footsteps on the stone slate ground.

RAPO output.

> a one-shot take of a man practicing Tai Chi sword in an antique courtyard. The man is dressed in a white shirt and black pants, and he is performing a series of movements with the Tai Chi sword. The video captures the beauty of the martial art and the practitioner's concentration. The background music is a soft melody played on a guzheng, which adds to the serene atmosphere. The video is 5 seconds long.

VPO output.

> A modern Chinese traditional style enthusiast practices Tai Chi sword in a serene antique courtyard bathed in soft morning light. The practitioner moves gracefully and continuously, performing fluid thrusts and elegant cleaves with the Tai Chi sword, showcasing the beauty and poise of the martial art. The intricate details of the courtyard's traditional architecture, such as ornate wooden carvings and arched windows, create a harmonious backdrop. The practitioner's focused expression and deliberate movements exude concentration, while the gentle light enhances the delicate interplay of shadows on the stone slate ground beneath their feet. The scene captures a moment of tranquility and elegance.

SCMaPR output.

> A single continuous shot begins in an antique courtyard bathed in soft morning light. A modern Chinese traditional style enthusiast practices Tai Chi sword alone, performing slow, fluid, and continuous movements. The sequence includes precise thrusting and cleaving actions with the sword, highlighting the aesthetic beauty of the martial art and the practitioner's deep concentration. The scene is accompanied by a soft guzheng melody, the sound of the sword blade cutting through the air, and light footsteps on the stone slate ground.

Prompt-A-Video (OpenSora) output.

> A one-shot video showing a Tai Chi sword practitioner in a charming antique courtyard garden, adorned with traditional Chinese elements, performing a continuous routine in morning light. The footage focuses on the elegant sword movements, including thrusting and cleaving actions, and the practitioner's focused concentration. A soft guzheng melody plays in the background, complemented by the sounds of the sword slicing through the air and footsteps on stone pavement. The video maintains a single steady shot throughout, with no additional actions or edits. (Duration: 5 seconds)

Prompt-A-Video (CogVideoX) output.

> In a serene, sun-kissed courtyard of ancient stones, a masterful practitioner dons a traditional silk robe, wielding a sword that dances through the misty morning. The slow, graceful movements of Tai Chi sword weave through the tranquil air,

highlighting the steel's gentle arc. The ensemble of robe, sword, and practiced form is a beauty to behold. The practitioner's focus is deep, as if the sword slices through more than just the gentle light. A soothing guzheng melody and the soft whoosh of the blade through the air complement the gentle tap of the sandals on the stone slate, capturing five serene seconds of a captivating martial art in harmony with the peaceful morning.

WanPE-4B output.

The video consists of 1 shot. This video is a one-shot take of a modern Chinese traditional style enthusiast practicing Tai Chi sword alone in an antique courtyard during the morning light. The scene is set in a modern Chinese traditional style antique courtyard. The overall visual style features a soft, clean photographic texture. The narrative perspective is objective, captured with a stable, professional shooting style and a slow, continuous pace. The audio consists of a gentle traditional music melody, the sound of the sword blade cutting through the air, and the sound of light footsteps on the stone slate ground.

Shot 1 [00.00-05.00s]. The scene opens with a full shot, framed at an eye-level angle with a deep depth of field that keeps all elements in focus, with a strong light-dark contrast presenting a mid-tone mid-key characteristic. The frame is a balanced composition, with the practitioner positioned in the center of the frame in the midground, and the courtyard's architecture creating a natural frame around him. The scene is set in a rectangular courtyard paved with large, flat gray stone slabs, enclosed by a dark brown wooden building with a dark gray tiled roof and a balcony with a intricate wooden lattice. A large, dark brown wooden door is centered on the back wall, and a smaller wooden window is on the right wall. A stone planter with a small green bush sits on the left, and a low wooden bench is on the right. The morning sun casts a warm, golden light across the scene, with the sunlight coming from the left, casting a warm, golden light on the subject and the courtyard, creating soft shadows. The man, an East Asian in his 30s-40s with a light skin tone, has a sharp jawline, a defined nose, and a focused expression with a slightly furrowed brow. He has short, neatly styled black hair. He is dressed in a loose-fitting, long-sleeved off-white shirt with a mandarin collar and a subtle, lighter off-white circular pattern, and wide-legged off-white trousers, and is barefoot. He holds 1 Tai Chi sword, which has a straight, slender, silver-colored metal blade, a dark brown wooden grip, and a dark brown, almost black, circular guard. The man practices Tai Chi with the sword. Standing in the center of the courtyard, he performs a series of slow, continuous, and fluid movements. He begins with the sword held in a ready position, then smoothly extends his arm forward in a thrusting motion, the sword moving gracefully through the air. The camera slowly moves forward, gradually tightening the frame from a full shot to a medium full shot, drawing the viewer's attention more closely to his form and concentration. His movements are synchronized with a soft, melodic guzheng background music, and the gentle swoosh of the sword blade cutting through the air, and the sound of light footsteps on the stone slate ground. He follows through with a cleaving motion, the sword arcing horizontally in front of him, his body turning slightly with the movement. His expression remains one of deep concentration throughout. Finally, he returns to a stable, centered stance, the sword held horizontally before him, his gaze fixed forward.

J.1.2 Latitude and Longitude Lecture

User request (Knowledge, 15 seconds, intent-level).

Generate a classroom lecture video for junior high school geography on the concept of latitude and longitude lines. The speech should be at a normal pace, clear, and easy to understand. A diagram of latitude and longitude lines is presented on a blackboard. The entire video is filmed in a single, static shot.

RAPO output.

a classroom lecture on the concept of latitude and longitude lines. The teacher is standing in front of a blackboard and drawing a diagram of the latitude and longitude lines. The video is of a normal pace, and the teacher's voice is clear and easy to understand.

VPO output.

A teacher stands at the front of a brightly lit classroom, writing on a blackboard covered in neatly drawn diagrams, including a detailed illustration of latitude and longitude lines. The teacher, dressed in professional attire, speaks clearly and confidently, their gestures animated but controlled, ensuring their words are easy to follow. The camera remains static, capturing the entire classroom scene, including rows of students seated attentively at their desks, taking notes. The teacher points to the diagram on the blackboard, emphasizing key points with their pen, explaining the concepts of latitude and longitude lines in a way that seems engaging and comprehensible for junior high school students. The classroom's warm tones and organized setting enhance the academic atmosphere.

SCMaPR output.

A junior high school geography teacher stands in a classroom, delivering a clear and easy-to-understand lecture on the concept of latitude and longitude lines. The teacher speaks at a normal, steady pace. A detailed diagram of latitude and longitude lines is clearly drawn on a large blackboard behind the teacher. The entire video is filmed in a single, static, wide-angle shot that captures the teacher and the blackboard.

Prompt-A-Video (OpenSora) output.

A 15-second video showing a static shot of a teacher conducting a lesson on latitude and longitude lines in a junior high school geography classroom. The teacher writes explanatory notes on a blackboard chart displaying diagrams of latitude and longitude lines. The footage maintains a single fixed frame, with the teacher presenting in a clear and detailed manner. The classroom environment provides context without hindering the viewing experience. The video focuses on the teacher and the educational materials on the board.

Prompt-A-Video (CogVideoX) output.

An intelligent instructor stands in a well-organized junior high school classroom, addressing a diagram of intricate latitude and longitude lines on a clean blackboard. The educator speaks clearly at a moderate pace, explaining the concepts of these lines in detail. The image is steady, capturing a wide angle of the classroom with natural light pouring in through the windows. The instructor is attired in professional clothing, and the board is well-supplied with examples. The video concludes with the teacher emphasizing a key point, summarized by the clear labels on the board. The 15-second clip is a concise and informative lesson in the geography of latitude and longitude.

WanPE-4B output.

The video consists of 1 shot. This junior high school geography lecture video on latitude and longitude lines is set in a modern classroom. A teacher explains the concept of latitude and longitude lines. The scene is a classroom. The overall visual style features a clean, sharp digital texture. The narrative is presented from an objective perspective, with a steady shooting style and a continuous, uncut pace, and the audio consists of a clear, educational speech.

Shot 1 [00.00-15.00s]. The scene opens with a static, eye-level medium shot with a deep depth of field, where the frame is a stable medium shot. The scene is set during the day, with the lighting dominated by a soft fluorescent tube light,

the frame presenting a mid-tone mid-key characteristic with a moderate light-dark contrast. The frame is composed with a 30s-40s East Asian female teacher with a light skin tone, a round face, and a short, dark brown bob, positioned in the left third of the frame, facing three-quarters profile to the right. She wears a dark navy blue long-sleeved button-down shirt and a patterned purple and white necktie. The majority of the frame is occupied by a large, rectangular black blackboard in the midground and background, which displays a diagram of latitude and longitude lines in white chalk, with a rectangular white chalk tray at its base. The teacher and the blackboard are in sharp focus. The scene is evenly lit by a soft fluorescent tube light, which forms a side light effect on the teacher, presenting a medium lighting ratio, while the blackboard is illuminated by a soft fluorescent tube light. The teacher is explaining the geographical concept. With a focused expression and a slightly open mouth, she stands in front of the blackboard and gestures with her right hand towards the diagram, her index finger and palm moving smoothly to trace the lines. She speaks clearly and continuously, her voice calm and instructional, with a medium pitch, a neutral and clean texture, and a clear, youthful style. Her pronunciation is very clear with natural liaison, and her speech rate is steady and moderate, with a natural rhythm and a flat-leaning intonation. She pauses naturally at logical sentence boundaries, maintaining a medium and steady volume. She says, 'latitude and longitude lines. These lines help us to locate any place on the Earth's surface. Latitude lines are parallel to the equator, while longitude lines run from the North Pole to the South Pole.' As she speaks, her hand movements are synchronized with her explanations, her hand moving horizontally to illustrate the parallel nature of latitude and then vertically to demonstrate the vertical run of longitude. She continues, 'These lines form a grid that divides the Earth into various regions.' Finally, she concludes her point, her hand resting on the blackboard as she maintains her gaze on the diagram and continues her explanation.

J.2 Scene-Level Requests

J.2.1 Barbershop Haircut Mishap

User request (Action, 15 seconds, scene-level).

In a small barbershop, there are only a customer and a barber. During the haircut, a series of accidents happen. The barber panics and hastily tries to fix it, but the more he tries, the worse it gets, resulting in an increasingly bizarre and unsalvageable haircut, which is comical and funny. Use multiple shots to show the changes in the barber's expression during the process. The video should be paired with lighthearted and quirky background music, creating a silly and relaxed atmosphere. The visual style should be a lifelike, documentary texture.

RAPO output.

a group of people in a barbershop, with one person sitting in a chair and another person cutting their hair. The haircut goes wrong, and the barber tries to fix it. The video is 15 seconds long, and it has a lighthearted and quirky tone.

VPO output.

A small, cozy barbershop with warm lighting and vintage decor features a barber intently working on a customer's haircut. Initially, the scene captures the barber's confident demeanor as he carefully trims the customer's hair. However, as the haircut progresses, a series of minor mishaps occur—dropped tools, accidental snips of hair flying, and uneven cuts. The barber's expression quickly shifts from calm focus to growing concern, then to sheer panic. Despite his efforts to fix the mistakes, each attempt worsens the haircut, leading to a comical and absurdly uneven style. The camera alternates between close-ups of his frustrated expressions and the evolving haircut, emphasizing the humor in the situation. The lighthearted

and quirky background music complements the scene, maintaining a playful and relaxed atmosphere.

SCMaPR output.

In a small, realistically lit barbershop, a single customer sits for a haircut. A series of minor accidents begins, shown through distinct shots. The barber's initial calm expression shifts to panic as he tries to fix each mistake. Each attempt worsens the haircut, escalating to a bizarre and unsalvageable style. The sequence of shots clearly shows the barber's increasingly frantic and comical expressions. The video has a lifelike, documentary texture with lighthearted, quirky background music, creating a silly and relaxed atmosphere.

Constraints: The barber gets panicked and hastily tries to fix the accidents.

Prompt-A-Video (OpenSora) output.

In a quaint barbershop, a customer sits in a chair, and a barber is giving him a haircut. The footage depicts a sequence of mishaps occurring during the cutting process, leading the flustered barber to attempt to salvage the haircut. His growing frustration mirrors the increasingly unusual and salvageable haircut. The video uses multiple angles to capture his emotional changes. Lighthearted, quirky music plays in the background, maintaining a playful and relaxed tone. A documentary-style visual approach enhances the humor. (15 seconds)

Prompt-A-Video (CogVideoX) output.

In a cozy, sunlit barbershop with leather chairs and polished reflections, a flustered barber attends to a patient client. As mishaps unfold, the barber's concerned expression deepens. He frantically tries to rectify the tangled mess, but it only escalates into a delightfully absurd haircut. The barber's face reflects his growing distress through a series of close-ups, from concern to exasperation. The lighthearted, quirky background music and the lifelike, documentary-style visuals capture the comical essence of the situation, playing out to a humorous conclusion in 15 seconds. The barber and client share a laugh, and the barber shrugs, 'Well, it's a unique look!'

WanPE-4B output.

The video consists of 5 shots. This video is a comical funny barbershop haircut mishap video, in which a barber's attempt to fix a haircut goes awry. The scene is set in a small barbershop. The overall visual style has a lifelike, documentary texture. The narrative is presented from an objective perspective with a stable shooting style and a moderate editing pace. The audio consists of a lighthearted, quirky background music track, which creates a silly and relaxed atmosphere.

Shot 1 [00.00-02.80s]. The scene opens with a medium shot at an eye-level angle, with a shallow depth of field that keeps the subjects in focus. The frame is a balanced composition, with a customer in the left foreground and a barber in the right midground, under a soft warm-yellow incandescent top light with a medium lighting ratio, and a cool-white fluorescent background light. The customer, a 30s-40s Asian man with a light tan skin tone, a round face, and short black hair, is seen in profile, facing right. He wears a dark brown collared shirt and a dark gray towel around his neck, and sits with a calm, neutral expression. The barber, a 30s-40s Asian man with a light tan skin tone, a round face, and short black hair, is seen in a 3/4 profile, facing left. He wears a black short-sleeved collared shirt, a black apron, and a dark gray towel around his neck. The barber is actively cutting the customer's hair, his expression focused. He holds a silver metal hair clipper, a handheld electric device, in his right hand, and a silver metal comb, a flat rectangular tool, in his left hand. The customer sits still, looking forward with a calm expression. The barber focuses on the haircut, his expression calm and concentrated, and carefully trims the customer's hair with the clipper, moving it

steadily from left to right. The camera slowly pushes in and pans right, tightening the frame on the action. A lighthearted, quirky background music track featuring a pizzicato string section and a light percussion with a steady pulse at a moderato tempo and mezzo forte intensity begins, establishing a playful mood.

Shot 2 [02.80-05.00s]. A hard cut shifts the perspective to a medium close-up of the barber at an eye-level angle, with a shallow depth of field. The frame is a right-weighted composition, with the barber in the center-right of the frame, seen in a 3/4 profile, facing left, under a soft warm-yellow incandescent top light with a medium lighting ratio, and a cool-white fluorescent background light. The customer is partially visible in the left foreground, with the back of his head and right shoulder in the frame. The barber's expression suddenly changes. His eyes widen and his mouth opens slightly as he looks down, his expression shifting from calm to a startled and panicked state. He abruptly stops the haircut. The barber's right hand, holding the silver metal hair clipper, moves away from the customer's head, and he quickly looks down at the clipper, his left hand still holding the silver metal comb. The customer remains still. The camera slowly moves forward and to the right, and slowly pans left, keeping the barber's face centered. The lighthearted background music continues unchanged.

Shot 3 [05.00-07.00s]. Following the barber's reaction, a hard cut presents a close-up of the customer's head from a slightly high angle, with a shallow depth of field. The frame is a center composition, with the customer's head in the center of the frame, seen from a 3/4 back view, under a soft warm-yellow incandescent top light with a medium lighting ratio, and a cool-white fluorescent background light. The customer's hair is now unevenly cut, with a noticeable bald patch. The customer remains still, his head tilted slightly down, his expression calm and indifferent. The barber's hand enters the frame from the right, holding the silver metal comb, and gently runs it across the customer's forehead. The camera slowly moves forward and to the left, and slowly pans right, exploring the result of the mistake. The lighthearted background music continues unchanged.

Shot 4 [07.00-10.00s]. A hard cut returns to a medium close-up of the barber at an eye-level angle, with a shallow depth of field. The frame is a right-weighted composition, with the barber in the center-right of the frame, seen in a 3/4 profile, facing left, under a soft warm-yellow incandescent top light with a medium lighting ratio, and a cool-white fluorescent background light. The customer is partially visible in the left foreground, with the back of his head and right shoulder in the frame. The barber attempts to fix the haircut. He furrows his brows, his eyes wide with tension, and his mouth tight, showing a flustered and anxious expression. He brings the silver metal hair clipper up to his mouth, as if to trim his own lip, then quickly moves it back to the customer's head, making a quick, sharp cutting motion. The customer remains still. The camera slowly moves forward and to the right, and slowly pans left, following the barber's frantic actions. The background music's rhythm becomes slightly more pronounced and urgent, with a repeated musical motif.

Shot 5 [10.00-15.00s]. A final hard cut reveals the outcome in a medium shot at an eye-level angle, with a shallow depth of field. The frame is a center composition, with the customer in the left foreground, seen in profile, facing right, and the barber in the right midground, seen in a 3/4 profile, facing left, under a soft warm-yellow incandescent top light with a medium lighting ratio, and a cool-white fluorescent background light. The customer's hair is now in a severely messed-up state. The customer looks down, his expression calm and indifferent. The barber assesses the damage. He looks at the customer's hair, his eyes widening in shock and his mouth falling open, his expression transitioning to complete astonishment and despair. He lowers the silver metal hair clipper, his hand hovering in the air, and then brings his left hand up to his own mouth, his face contorted in a expression of defeat. The customer remains still. The camera slowly moves forward and to the right, and slowly pans left, focusing on the barber's final reaction. The background music reaches a crescendo and then fades out.

J.3 SHOT-LEVEL REQUESTS

J.3.1 LATE-NIGHT EXAM PREPARATION

User request (Speech, 15 seconds, shot-level).

> Generate a dialogue scene in a study room, focusing on capturing the emotional performance.
>
> Shot 1: Two-shot, mid shot. Late at night, with open notebooks and two cups of steaming coffee on the table. The boy throws down his pen, leans back in his chair, and says with a self-deprecating laugh of resignation: 'I'm done. I'll just wing the exam tomorrow.' The girl looks up from her pile of books and glances at him.
>
> Shot 2: Cut to a close-up of the girl. She pushes her notes towards him, her tone urgent and sharp: 'If you fail again, you won't have enough credits.' Shot 3: Cut to a close-up of the boy. The boy slams the table and takes a deep breath. He picks up his pen again, flips a page in his notebook, and says slowly, as if psyching himself up: 'Alright, one more time.' The girl hums softly off-screen. Just as a self-mocking smile appears on his lips, he flips to the next page, and his smile suddenly freezes. His voice instantly shifts from determined to a wail: 'Wait! This chapter... the professor said it wouldn't be on the test, and I spent three whole days memorizing it!' Shot 4: Cut to a close-up of the girl. She is stunned for a moment, then bursts out laughing. Halfway through her laugh, she frowns, her tone shifting from amused to sympathetic and helpless: 'Then you might as well study this part instead.' She flips open her notebook and pushes it towards him, a smile still lingering on her face.

RAPO output.

> a dialogue scene between two students in a study room. The boy is sitting at a desk and the girl is sitting next to him. The boy is trying to study for an exam while the girl is looking at him. The girl then pushes her notes towards the boy and says "If you fail again, you won't have enough credits." The boy then slams the table and takes a deep breath. The girl then pushes her notes towards him again and says "Then you might as well study this part instead." The video ends with the boy sitting at the table with a smile on his face.

VPO output.

> A dimly lit study room late at night, with the warm glow of desk lamps illuminating stacks of open notebooks and a steaming cup of coffee on the wooden table. The boy leans back in his chair with a self-deprecating laugh, his shoulders sagging as he declares he will wing the exam. The girl, seated across from him, looks up from her books, her expression a mix of concern and determination. She pushes her notes toward him, her urgent words emphasizing the stakes of failure. The boy's demeanor shifts dramatically as he picks up his pen, determined to study again, only to slump in despair when he realizes he has wasted time on an irrelevant chapter. His expression freezes as he looks at the page, defeated. The girl bursts into laughter before softening into a sympathetic smile. She flips open her notebook and pushes it toward him, her expression a mix of amusement and encouragement.

SCMaPR output.

> Shot 1: A mid shot, two-shot, late at night in a study room. Open notebooks and two steaming cups of coffee are on a table. A boy throws down his pen, leans back in his chair, and says with a self-deprecating laugh of resignation, 'I'm done. I'll just wing the exam tomorrow.' The girl looks up from her pile of books and glances at him.

> Shot 2: Cut to a close-up of the girl. She pushes her notes towards him, her tone urgent and sharp, 'If you fail again, you won't have enough credits.' Shot 3: Cut to a close-up of the boy. He slams the table, takes a deep breath, picks up his pen again, and flips a page in his notebook. He says slowly, as if psyching himself up, 'Alright, one more time.' The girl hums softly off-screen. A self-mocking smile appears on his lips, but as he flips to the next page, his smile suddenly freezes. His voice shifts from determined to a wail, 'Wait! This chapter... the professor said it wouldn't be on the test, and I spent three whole days memorizing it!' Shot 4: Cut to a close-up of the girl. She is stunned for a moment, then bursts out laughing. Halfway through her laugh, she frowns, her tone shifting from amused to sympathetic and helpless, 'Then you might as well study this part instead.' She flips open her notebook and pushes it towards him, a smile still lingering on her face.

Prompt-A-Video (OpenSora) output.

> A study room video clip depicting a young boy and girl during a late-night study session. The table has open textbooks and steaming coffee cups. The boy, seated in a chair, expresses resignation after giving up on studying, to which the girl responds from across the table. The video then cuts to a close-up of the girl, who persuades the boy to continue studying. The boy comes to realize he invested effort in covering material that won't be on the exam, leading to his frustration. The girl comforts him with a suggestion to focus on that part instead. The video captures their emotional exchange in a sympathetic light. (15 seconds)

Prompt-A-Video (CogVideoX) output.

> In a dimly lit study room, a young man and woman engage in a poignant conversation over steaming cups of coffee and scattered notes. The man, with a resigned chuckle, admits he's given up, planning to wing the exam. The woman, concerned, urges him to reconsider, pointing out the importance of not failing again. The man, with newfound determination, returns to his studies, but just as he's about to begin, a look of despair washes over him as he realizes he's wasted time on an unused chapter. The woman, sympathetic, can't help but laugh at the absurdity, then offers guidance, suggesting they focus on the more relevant material. The scene captures the essence of their emotional exchange through a mix of mid shots and close-ups, showcasing the intimacy of their late-night study session.

WanPE-4B output.

> The video consists of 4 shots. This film scene, a casual study room dialogue scene of a young student and his female companion, depicts a young man in a study room giving up on his exam, only to have a unexpected realization. The scene is set in a study room. The overall visual style is cinematic, with a soft, naturalistic texture. The narrative is presented from a third-person objective perspective, using a stable camera and a moderate editing pace that follows the conversation. The audio consists of dialogue and subtle sound effects.
>
> Shot 1 [00.00-02.83s]. A static, eye-level medium shot establishes the scene with a balanced two-shot composition. The scene is set at night, with the lighting dominated by a soft incandescent light, presenting a low-key characteristic with a moderate light-dark contrast. The frame has a soft, naturalistic texture, with a low color saturation overlaid with a weak warm-toned grading. The frame is filled with a late-night study scene, with a wooden table in the midground. On the table, there are multiple open paper notebooks with white pages and handwritten text, two white ceramic coffee cups with saucers emitting a gentle, rising wisps of white steam, and a black plastic or metal pen. In the background, a large window reveals the darkness of the night, and a bookshelf is partially visible on the right. On the left side of the frame, a late teens to early twenties young man of mixed East Asian and Caucasian descent with a light tan skin tone, a sharp jawline, and a defined nose, is positioned in a 3/4 profile facing right. He has short, styled black hair and

wears a heather gray long-sleeved shirt with a black-and-white geometric pattern on the sleeves, and dark gray pants. On the right side of the frame, a late teens to early twenties young woman of Caucasian descent with a fair skin tone, a round face, and a straight nose, is positioned in a 3/4 profile facing left. She has shoulder-length brown hair with a bangs, tied back in a ponytail, and wears a pale yellow t-shirt under a dark olive green cardigan, and dark blue jeans. The soft incandescent light forms a front-side light effect on the young man and the young woman, presenting a medium lighting ratio. The young man, with a frustrated expression, abruptly throws his black pen down on the table, the action accompanied by a soft thud. He then leans back in his chair, his body relaxing, and looks at the young woman with a self-deprecating laugh, his expression shifting to one of resignation. He says with a medium speech rate and a slightly flat rhythm, his volume medium, with a slight sighing feel and a faint laughing tone, in a resigned tone, 'I'm done. I'll just wing the exam tomorrow.' The young woman, who was looking down at her pile of books, looks up at him, her expression neutral, and glances at him.

Shot 2 [02.83-05.00s]. The man's declaration of giving up prompts a hard cut to a close-up of the young woman, a static, eye-level shot with a shallow depth of field that isolates her against a blurred background. The frame is a close-up, with the young woman positioned in the center of the frame, facing 3/4 left. The young man's shoulder is visible in the foreground on the left. The young woman's face is in sharp focus, her brown hair framing her face. The young woman looks intently at the young man, her expression serious. She pushes her notes towards him, and speaks to him with a serious expression, her eyes wide. She speaks with a fast-leaning speech rate and a steady rhythm, her volume medium, with a firm and urgent intonation, in a sharp, urgent tone, 'If you fail again, you won't have enough credits.'

Shot 3 [05.00-11.00s]. A hard cut reverses the angle, returning to a close-up of the young man, a static, eye-level shot with a shallow depth of field. The frame is a close-up, with the young man positioned in the center of the frame, facing 3/4 right. The young woman's shoulder is visible in the foreground on the right. The young man's face is in sharp focus, his black hair styled, and his expression initially one of frustration. The young man slams the table with a soft thud, and takes a deep breath. He then picks up his black pen again, and flips a page in his notebook, the sound of a crisp page turn is heard. He says slowly, as if psyching himself up, with a medium speech rate and a steady rhythm, 'Alright, one more time.' The young woman hums softly off-screen. Just as a self-mocking smile appears on his lips, he flips to the next page, and his smile suddenly freezes. His voice instantly shifts from determined to a wail, his speech rate accelerates, his volume increases, and his intonation rises sharply, with a strong sense of exasperation, as he exclaims, 'Wait! This chapter... the professor said it wouldn't be on the test, and I spent three whole days memorizing it!'

Shot 4 [11.00-15.00s]. The man's exclamation triggers a hard cut back to a close-up of the young woman, a static, eye-level shot with a shallow depth of field. The frame is a close-up, with the young woman positioned in the center of the frame, facing 3/4 left. The young man's shoulder is visible in the foreground on the left. The young woman's face is in sharp focus, her expression initially one of shock. She is stunned for a moment, then bursts out laughing. Halfway through her laugh, she frowns, her tone shifting from amused to sympathetic and helpless. She speaks with a medium speech rate, her volume medium, with a varied rhythm, and a tone that transitions from a clear laugh to a softer, sympathetic tone, in a playful tone, 'Then you might as well study this part instead.' She flips open her notebook and pushes it towards him, a smile still lingering on her face.