WANPEBENCH: TOWARDS UNIFIED EVALUATION OF PROMPT ENHANCERS FOR VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern video generators can realize increasingly complex visual narratives, positioning the prompt enhancer (PE) as a critical bridge from concise user instructions and multimodal references to structured cinematic plans. However, existing PE evaluation relies on rendered videos, imposing substantial computational and human costs, slowing PE training and iteration, and conflating PE quality with downstream generator behavior. To address this gap, we introduce WanPEBench, the first unified benchmark for direct PE evaluation across text-to-video, image-to-video, and reference-to-video prompt enhancement. It comprises 1,100 expert-verified cases and 1,005 visual assets, spanning 35 fine-grained tasks with diverse temporal, cinematic, audiovisual, and multi-reference requirements. In addition, we develop WanPEBench evaluation, an evidence-grounded framework that combines modality-aware fact extraction with rubric-based assessment across 24 criteria. Our systematic evaluation of representative open- and closed-source PE methods reveals an emerging shift from fine-grained descriptive expansion toward intent-preserving cinematic planning, while the caption-reconstruction and forward-refinement methods show complementary strengths in cinematic coverage and semantic fidelity or internal coherence, respectively. Human validation shows that WanPEBench scores align closely with expert judgments of enhanced prompts and downstream videos from Wan3.0 and MiniMax-H3, indicating that prompt-level evaluation reliably reflects downstream utility.

1 INTRODUCTION

Rapid advances in large-scale video generation models (Wan Team, 2026; Seed, 2026; MiniMax Team, 2026) have transformed content creation by enabling diverse video synthesis from natural-language instructions and multimodal references. Despite their remarkable capabilities, the quality and fidelity of generated videos remain strongly influenced by input conditions that adequately express user intent and generation requirements (Wang et al., 2025b; Zhu et al., 2026). Concise or ambiguous user instructions can lead generators to misinterpret the bindings between subjects and actions, the intended event order, or the roles of visual references, revealing a mismatch between user intent and model interpretation. A prompt enhancer (PE) addresses this mismatch by converting user instructions and conditioning inputs into an enhanced prompt that guides downstream generation, making PE quality an important factor in high-quality video synthesis (Zhu et al., 2026).

However, PE development remains constrained by an indirect evaluation loop: validating each update requires downstream video synthesis followed by automatic scoring or blind human comparisons (Zhu et al., 2026; Cheng et al., 2025; Yang et al., 2026; Gao et al., 2025a; Ji et al., 2024; Gao et al., 2025b). These computational and human costs delay feedback for training and refinement. Beyond this inefficiency, rendered videos reflect the joint influence of PE quality and generator capability, making observed errors difficult to attribute reliably to either component. These limitations motivate direct evaluation of enhanced prompts to assess PE quality independently of generator behavior and provide rapid, PE-specific feedback throughout iterative refinement.

In this paper, we introduce **WanPEBench**, the first unified benchmark for direct evaluation of video prompt enhancers. As illustrated in Figure 1(a), WanPEBench spans T2V-PE, I2V-PE, and R2V-PE, covering text-only, first-frame, and multi-reference conditioning. Figure 1(b) summarizes 1,100 expert-verified cases and 1,005 visual assets, covering 35 fine-grained task types with broad varia-

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.

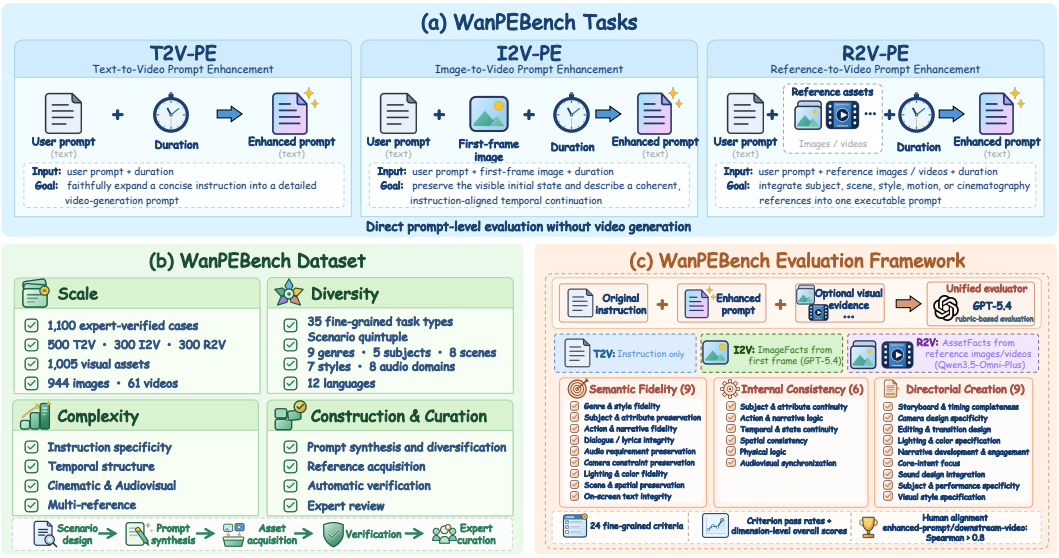


Figure 1: **Overview of WanPEBench.** (a) WanPEBench unifies T2V-PE, I2V-PE, and R2V-PE for direct prompt-level evaluation without video generation. (b) The dataset contains 1,100 expert-verified cases and 1,005 visual assets, covering diverse tasks, content, and conditioning complexity through a four-stage construction pipeline. (c) The evidence-grounded framework combines modality-aware evidence construction with rubric-based evaluation across 24 fine-grained criteria organized into Semantic Fidelity, Internal Consistency, and Directorial Creation.

tion in content, language, and conditioning complexity. To ensure diversity and quality, we employ a staged dataset construction pipeline combining broad scenario coverage with rigorous verification. Beyond dataset construction, the evidence-grounded framework in Figure 1(c) evaluates each enhanced prompt against the original instruction and modality-aware facts extracted from visual inputs where applicable. Its 24 fine-grained criteria form three complementary dimensions: Semantic Fidelity verifies the preservation of source requirements, Internal Consistency examines narrative, spatiotemporal, and audiovisual coherence, and Directorial Creation assesses the quality of creative guidance for realizing the intended video. A unified rubric-based evaluator reports criterion pass rates and dimension scores, yielding an interpretable profile of PE quality.

Using WanPEBench, we systematically evaluate representative open- and closed-source PE systems across T2V-PE, I2V-PE, and R2V-PE, revealing four key findings, most notably an emerging shift toward cinematic planning among industry-developed PE and observed links between specific training strategies and capability profiles. (i) Prompt enhancement is evolving from fine-grained descriptive expansion toward intent-preserving cinematic planning as video generators become increasingly capable. (ii) Caption-reconstruction methods achieve broader cinematic coverage, whereas forward-refinement methods favor semantic preservation or internal coherence. (iii) Increasing temporal and reference complexity generally degrades PE performance. (iv) WanPEBench evaluation scores achieve Spearman correlations above 0.8 with expert rankings of enhanced prompts and downstream videos produced by Wan3.0 and MiniMax-H3, indicating that prompt-level evaluation reliably reflects downstream utility across generators.

Our contributions are summarized as follows:

- We introduce WanPEBench, the first unified benchmark for direct evaluation of video prompt enhancers across T2V-PE, I2V-PE, and R2V-PE, comprising 1,100 expert-verified cases and 1,005 visual conditioning assets spanning 35 fine-grained task types.
- We propose WanPEBench Evaluation, an evidence-grounded framework that combines modality-aware fact extraction for visually conditioned tasks with rubric-based assessment.
- We systematically evaluate representative PE, revealing a shift toward cinematic planning and capability differences associated with training strategies, while confirming human alignment and downstream utility across generators.

Table 1: **Comparison with existing video-generation benchmarks. Cine.**: cinematic and camera control; **Scripted**: temporal or shot-level prompts; **Audio Req.**: explicit audio requirements in prompts; **Multi-lang.**: three or more languages; **Video Ref.**: conditioning on reference videos; **Direct PE**: direct prompt-level PE evaluation decoupled from downstream video generators.

Benchmark	T2V	I2V	R2V	Cine.	Scripted	Multi-shot	Audio Req.	Multi-lang.	Video Ref.	Direct PE	# Metrics	# Prompts	# Ref. Assets
VBench (Huang et al., 2023)	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	16	∼1,600	–
EvalCrafter (Liu et al., 2024)	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	17	700	–
Video-Bench (Han et al., 2025)	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	9	419	–
OpenS2V-Nexus (Yuan et al., 2025)	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	6	180	180
UniVBench (Wei et al., 2026a)	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗	21	200	864
MSAVBench (Wei et al., 2026b)	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗	20	286	165
StreamAV-Bench (Liu et al., 2026)	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	32	320	–
WanPEBench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**24**	**1,100**	**1,005**

2 Related work

Prompt Enhancement for Video Generation. As video generators increasingly support temporally structured and multimodally conditioned content (Wan Team, 2026; Seed, 2026; AI, 2026), prompt enhancers (PE) have become important for translating concise user intent into detailed instructions for downstream video generation. Existing approaches (Wang et al., 2025a; Jia et al., 2026; Long et al., 2025) broadly follow training-based or inference-time paradigms. Training-based methods learn prompt rewriting from constructed supervision or preference signals. VPO (Cheng et al., 2025) and Prompt-A-Video (Ji et al., 2024) optimize LLM-based rewriters using supervised and preference objectives. WanPE (Zhu et al., 2026) instead reconstructs detailed video captions from compressed prompts using video-grounded inverse pairs, while LingBot-Video-PE (Ma et al., 2026) maps expanded prompts into model-specific structured captions. In contrast, inference-time methods adopt diverse strategies: RAPO (Gao et al., 2025a) retrieves relevant examples, while SCMaPR (Yang et al., 2026) performs agentic verification and correction. LTX-2.5-PE (Lightricks, 2026) directly rewrites prompts using its integrated Gemma 4 text encoder (Team et al., 2026), whereas H3-Context-IR (MiniMax Team, 2026) employs a multi-stage multimodal workflow.

Benchmarks for Video Generation and Prompt Enhancement. Video-generation benchmarks have evolved from assessing T2V quality and prompt alignment (Huang et al., 2023; Liu et al., 2024; Han et al., 2025) to covering reference-conditioned, multi-shot, audio-video, and streaming generation (Yuan et al., 2025; Wei et al., 2026a;b; Liu et al., 2026). However, these benchmarks are designed to assess rendered videos rather than enhanced prompts. Applying them to PE evaluation requires an additional generation stage, increasing evaluation cost and latency while entangling PE performance with downstream generator behavior. As summarized in Table 1, WanPEBench instead evaluates enhanced prompts directly across T2V-PE, I2V-PE, and R2V-PE using standardized rubrics covering 24 criteria, enabling fine-grained assessment independent of downstream video generation while offering broad coverage of prompt structures and conditioning modalities at scale.

3 WanPEBench

3.1 Task Definition

A video prompt enhancer takes a user prompt, optional reference assets $\mathcal{R}$, and a target duration as input and produces an enhanced prompt for a downstream video generator, as illustrated in Figure 1(a). WanPEBench evaluates this output directly without requiring video generation, thereby separating PE quality from downstream generator behavior. Specifically, it assesses whether the enhanced prompt preserves user intent and conditioning constraints, remains internally consistent, and provides sufficient detail to specify the intended video.

To provide comprehensive coverage, WanPEBench introduces a unified benchmark for evaluating prompt enhancers across three distinct task settings:

- **Text-to-Video Prompt Enhancement (T2V-PE).** Given a user prompt and a target duration, with no reference assets ($\mathcal{R} = \varnothing$), the PE should produce a detailed prompt that faithfully expands the requested video content.

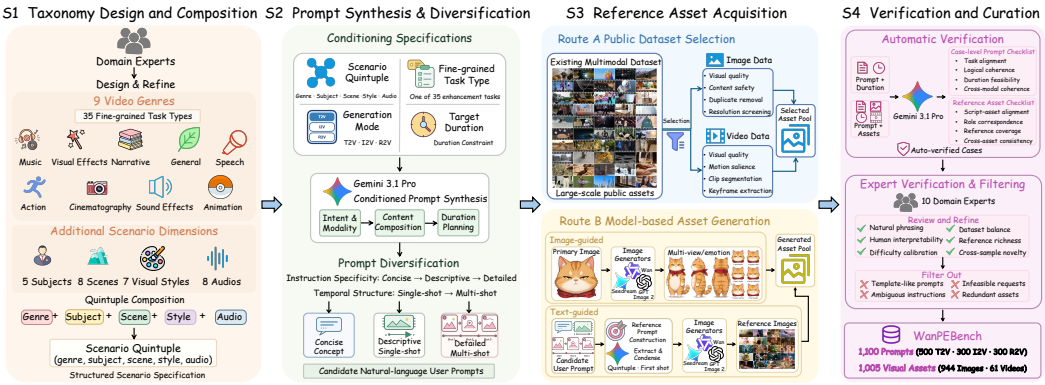


Figure 2: **WanPEBench data construction pipeline.** The four-stage pipeline comprises taxonomy design and composition, prompt synthesis and diversification, reference asset acquisition from public datasets and generative models, and automated verification followed by expert curation.

- **Image-to-Video Prompt Enhancement (I2V-PE).** Given a user prompt, a first-frame image ($\mathcal{R} = \{I_0\}$), and a target duration, the PE should preserve the visible initial state and describe a coherent, instruction-aligned temporal continuation.
- **Reference-to-Video Prompt Enhancement (R2V-PE).** Given a user prompt, a set of reference images or videos ($\mathcal{R} = \{a_i\}_{i=1}^{N}$), and a target duration, the PE should incorporate each asset according to its specified role, such as subject, scene, style, motion, or cinematography, into a single coherent generation prompt.

3.2 Data Construction

To construct a diverse and reliable benchmark, we develop a four-stage pipeline that combines automated synthesis and verification with expert-driven design and curation, as illustrated in Figure 2.

Stage 1: Taxonomy design and composition. Domain experts define nine video genres and refine them into 35 fine-grained task types to establish broad and structured task coverage. They further specify five subject categories, eight scene domains, seven visual styles, and eight audio domains to guide scenario composition. Using the genre and these four additional dimensions, each scenario is represented as a $(\text{genre}, \text{subject}, \text{scene}, \text{style}, \text{audio})$ quintuple.

Stage 2: Prompt synthesis and diversification. We sample a scenario quintuple and associate it with a generation mode (T2V, I2V, or R2V), a fine-grained task type, and a target duration. Gemini 3.1 Pro (Google, 2026) then synthesizes a user prompt from these specifications, jointly considering intent and modality, content composition, and duration planning. To reflect real-world prompting behavior, we vary instruction specificity from concise to detailed and temporal structure between single- and multi-shot formats, producing inputs ranging from short concepts to structured scripts.

Stage 3: Reference asset acquisition. We construct a diverse reference pool through two complementary routes. Public images and videos from existing datasets (Liu et al., 2025; Nan et al., 2025; Yuan et al., 2025; Liang et al., 2026) undergo modality-specific curation for quality, safety, and reference suitability. Model-generated references produced with Wan-Image (Mao et al., 2026), Seedream (Seedream et al., 2025), and GPT Image 2 (OpenAI, 2026b) follow image-guided and text-guided pipelines to create identity-consistent variants and scenario-aligned assets. The resulting pool supports single-image, multi-image, video-only, and hybrid image-video conditioning.

Stage 4: Verification and curation. Each candidate undergoes a two-stage quality-control process combining scalable model-based screening with expert review. Gemini 3.1 Pro first derives case-specific checks from the instruction and target duration to detect infeasible or inconsistent specifications. For reference-conditioned cases, it jointly inspects the instruction and candidate assets to verify role correspondence, script-asset alignment, reference coverage, and cross-asset consistency. Ten domain experts then refine the verified cases for clarity and difficulty, removing ambiguous, template-like, infeasible, or redundant content while preserving dataset balance and diversity.

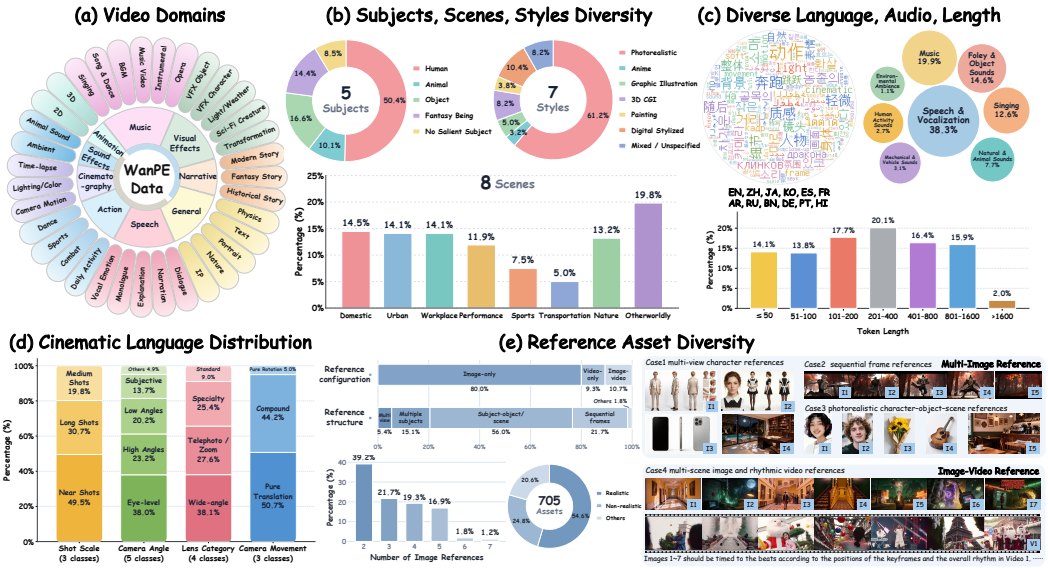


Figure 3: **Detailed Data Statistics and Diversity of WanPEBench.** The benchmark covers a broad taxonomy of video-generation domains and fine-grained tasks (a), diverse subjects, scenes, and visual styles (b), multilingual prompts, audio conditions, and prompt lengths (c), rich cinematic requirements (d), and diverse reference assets (e). More statistics are in Appendix A.3.

3.3 Data Statistics

Scale and Diversity. WanPEBench comprises 1,100 prompts across three PE tasks: T2V-PE (500), I2V-PE (300), and R2V-PE (300). As shown in Figure 3(a,b), its taxonomy spans 9 video genres and 35 fine-grained task types, while its scenarios cover 5 subject categories (*e.g.*, human and fantasy beings), 8 scene domains (*e.g.*, domestic and nature), and 7 visual styles (*e.g.*, photorealistic and 3D CGI), providing broad task, semantic, environmental, and aesthetic diversity.

Linguistic and Audio Diversity. In Figure 3(c), WanPEBench spans 12 languages (*e.g.*, English and Chinese) and 8 audio domains (*e.g.*, speech and vocalization, and music). Prompt lengths range from 6 to 4,120 tokens, encompassing concise instructions and long, temporally structured scripts.

Temporal and Cinematic Complexity. WanPEBench captures temporal and cinematic complexity through structured prompts containing 1–13 shots and rich cinematic language. As shown in Figure 3(d), it covers 3 shot scales (*e.g.*, near and long shots), 5 camera angles (*e.g.*, eye-level and low angles), 4 lens types (*e.g.*, wide-angle and telephoto/zoom), and 3 camera-movement classes (*e.g.*, pure translation and compound movement). Beyond these dimensions, the prompts also specify diverse composition patterns, transitions, lighting, and color schemes.

Reference Asset Diversity. As shown in Figure 3(e), WanPEBench contains 1,005 visual conditioning assets, with 944 images and 61 videos. I2V-PE uses 300 first-frame images, while R2V-PE includes 644 images and 61 videos in image-only, video-only, and hybrid configurations, with up to 7 images or 2 videos per case. Beyond modality diversity, R2V-PE covers multi-reference structures (*e.g.*, multi-view and multiple subjects) across realistic (54.6%), non-realistic (24.8%), and mixed or unspecified (20.6%) styles. Four representative cases illustrate multi-view characters, stylized frame sequences, photorealistic character–object–scene compositions, and multi-scene images paired with rhythm-guiding videos, supporting compositional, sequential, and cross-modal evaluation.

3.4 Evaluation Framework

Evaluating prompt enhancers through generated videos conflates prompt quality with downstream generator behavior while incurring substantial computational cost and feedback latency, slowing training and iteration. WanPEBench thus directly evaluates enhanced prompts along three dimensions: Semantic Fidelity (*SF*), Internal Consistency (*IC*), and Directorial Creation (*DC*). These di-

mensions respectively assess fidelity to the inputs, internal coherence, and creative directorial planning for realizing the intended video. Figure 1(c) summarizes the 24 fine-grained criteria.

Semantic Fidelity. *SF* assesses whether the enhanced prompt preserves the original instruction and relevant information from visual references. Genre and Style Fidelity (*GSF*) measures retention of genre, medium, and visual style. Subject and Attribute Preservation (*SAP*) examines subject identity, appearance, and attribute bindings, while Action and Narrative Fidelity (*ANF*) covers actions, event order, and narrative intent. Dialogue and Lyrics Integrity (*DLI*) verifies verbal content, language, and speaker attribution, while Audio Requirement Preservation (*ARP*) covers speech, music, and sound effects. Camera Constraint Preservation (*CCP*), Lighting and Color Fidelity (*LCF*), Scene and Spatial Preservation (*SSP*), and On-Screen Text Integrity (*OTI*) respectively assess camera constraints, visual conditions, spatial relations, and visible text. In I2V-PE, the first frame anchors the initial state. In R2V-PE, each reference's instruction-specified role determines the asset facts evaluated.

Internal Consistency. Whereas *SF* measures fidelity to the inputs, *IC* evaluates the logical and spatiotemporal coherence of the enhanced prompt. Subject and Attribute Continuity (*SAC*) assesses the stability of subject identity and attributes. Action and Narrative Logic (*ANL*) evaluates event progression, causal relations, and subject interactions, while Temporal and State Continuity (*TSC*) examines event order and state transitions. Spatial Consistency (*SPC*) assesses locations, relative positions, and movement paths. Physical Logic (*PL*) checks whether events follow real-world physics or established fictional logic. Audiovisual Synchronization (*AVS*) evaluates the alignment between visual events and associated dialogue, music, and sound effects.

Directorial Creation. *DC* assesses whether the enhanced prompt provides relevant, explicit, and creative direction to specify the intended video across temporal, cinematic, narrative, and audiovisual aspects. Storyboard and Timing Completeness (*STC*) evaluates shot organization and duration allocation. Camera Design Specificity (*CDS*) examines framing, viewpoint, focus, and camera movement, while Editing and Transition Design (*ETD*) evaluates shot connections and editing logic. Lighting and Color Specification (*LCS*) assesses illumination and color planning. Narrative Development and Engagement (*NDE*) measures content progression and variation, while Core-Intent Focus (*CIF*) checks whether added content supports the original request. Sound Design Integration (*SDI*), Subject and Performance Specificity (*SPS*), and Visual Style Specification (*VSS*) respectively evaluate audio design, subject presentation, and visual-style articulation.

Evidence-Grounded Fine-Grained Evaluation. We formulate evaluation as a two-stage process combining modality-aware evidence construction with unified criterion-level assessment. First, we construct this evidence using modality-specific extraction checklists. GPT-5.4 (OpenAI, 2026a) extracts static visual facts from image-only inputs, while Qwen3.5-Omni-Plus (Team, 2026) jointly analyzes all assets when video references are present, capturing static, temporal, audiovisual, and cross-asset evidence. Each fact is grounded in its source asset. The resulting *ImageFacts* for I2V-PE and *AssetFacts* for R2V-PE are constructed once and shared across methods. Second, GPT-5.4 serves as the unified evaluator across all three PE tasks. Given the original instruction, enhanced prompt, and applicable visual evidence, it assesses all 24 criteria using fixed criterion-specific rubrics and structured checklists. Shared evidence and fixed rubrics across methods enable consistent, interpretable criterion-level comparisons within each task setting. See Appendix B for more details.

4 Experiments

4.1 Experimental Setup

We evaluate ten PE systems: H3-Context-IR (MiniMax Team, 2026), WanPE-397B and WanPE-35B (Zhu et al., 2026), LTX-2.5-PE (Lightricks, 2026), LingBot-Video-PE (Ma et al., 2026), SCMaPR (Yang et al., 2026), VPO (Cheng et al., 2025), Prompt-A-Video-Cog and Prompt-A-Video-OS (Ji et al., 2024), and RAPO (Gao et al., 2025a). All systems are evaluated on T2V-PE, with multimodal-capable systems also evaluated on I2V-PE and R2V-PE.

4.2 Main Results

Table 2 reports PE performance across all three tasks. On T2V-PE, different methods lead in three dimensions, while WanPE-397B presents a balanced profile by ranking first in Directorial Creation

Table 2: **Main results on WanPEBench.** Fine-grained metrics report criterion-level pass rates, defined as the percentage of cases without corresponding issues. The *Overall* scores aggregate criterion-level issues according to their number and severity in each case within each dimension. The best and second-best results for each task are highlighted in **bold** and underlined, respectively.

Method	Semantic Fidelity										Internal Consistency						
	GSF ↑	SAP ↑	ANF ↑	DLI ↑	ARP ↑	CCP ↑	LCF ↑	SSP ↑	OTI ↑	***Overall*** ↑	SAC ↑	ANL ↑	TSC ↑	SPC ↑	PL ↑	AVS ↑	***Overall*** ↑
T2V-PE																	
WanPE-397B	89.20	70.00	80.80	**94.20**	62.00	57.40	80.40	80.60	94.80	74.89	91.20	87.40	94.60	82.40	92.80	94.40	91.51
WanPE-35B	86.40	71.80	82.60	88.00	59.60	54.40	73.20	81.20	**97.00**	72.83	91.60	89.20	92.40	81.60	91.60	90.20	91.40
H3-Context-IR	81.40	66.40	70.40	93.00	58.40	49.20	76.80	81.20	95.40	67.34	**97.20**	**95.40**	96.80	**98.00**	91.80	90.80	94.50
SCMaPR	**94.60**	**87.60**	**85.60**	80.20	**82.00**	**87.20**	**90.60**	**92.60**	95.00	**87.92**	94.60	92.60	96.80	91.40	93.60	**95.60**	95.06
LTX-2.5-PE	83.00	78.00	76.20	80.40	75.20	43.00	72.20	82.20	91.20	71.21	88.60	80.60	85.60	77.60	89.80	91.00	87.52
LingBot-Video-PE	67.80	28.00	30.40	72.00	54.00	16.80	28.00	45.40	84.60	41.61	91.20	91.00	93.80	66.60	92.40	93.20	90.96
VPO	74.00	65.00	45.60	76.20	63.40	45.00	66.00	66.40	89.00	62.80	92.60	90.00	**97.20**	91.40	93.00	94.00	**95.07**
Prompt-A-Video-Cog	84.60	47.60	35.40	78.60	69.20	29.00	58.60	44.40	91.20	56.48	82.40	81.80	93.40	86.20	89.40	89.80	91.09
Prompt-A-Video-OS	81.20	46.60	30.40	76.80	67.20	33.20	56.20	48.00	89.80	54.83	86.00	85.80	96.40	90.00	93.40	93.00	93.87
RAPO	37.20	6.40	1.00	72.60	52.40	24.20	25.40	4.80	79.00	15.47	80.80	64.00	88.20	76.80	**93.80**	93.80	88.30
I2V-PE																	
H3-Context-IR	64.67	**56.67**	67.00	**89.00**	47.33	**52.00**	**87.67**	**76.33**	**81.00**	**58.59**	**97.67**	**96.33**	**97.33**	**99.33**	89.00	94.33	**95.05**
LTX-2.5-PE	**84.67**	56.33	**71.00**	81.67	**65.00**	24.00	78.00	72.00	71.67	56.77	85.67	80.00	88.00	76.33	86.33	90.33	87.36
LingBot-Video-PE	72.67	54.33	23.33	77.00	57.33	10.67	42.33	73.00	70.67	45.70	95.67	**96.33**	90.00	66.67	**89.67**	**95.00**	91.14
R2V-PE																	
H3-Context-IR	**73.33**	**54.00**	**65.00**	**83.33**	56.67	**51.00**	**76.33**	**73.33**	**82.00**	**58.11**	**87.33**	**90.00**	**98.67**	**95.33**	**94.67**	**91.67**	**94.90**
LTX-2.5-PE	61.67	43.67	62.67	70.00	**71.00**	27.67	64.00	58.00	75.67	49.03	76.33	79.33	88.67	74.67	93.67	82.67	86.66

Method	Directorial Creation									
	STC ↑	CDS ↑	ETD ↑	LCS ↑	NDE ↑	CIF ↑	SDI ↑	SPS ↑	VSS ↑	***Overall*** ↑
T2V-PE										
WanPE-397B	**66.60**	81.00	63.20	**80.20**	68.60	91.20	77.60	81.60	**82.20**	**78.97**
WanPE-35B	59.80	80.20	**63.60**	75.60	66.60	94.20	83.00	80.40	78.60	78.15
H3-Context-IR	12.20	**89.00**	55.40	76.80	**76.40**	94.00	**93.60**	**86.40**	73.60	76.73
SCMaPR	6.20	55.20	41.60	48.60	61.00	**99.20**	13.20	74.80	70.60	49.18
LTX-2.5-PE	2.40	76.60	36.40	66.80	65.40	94.20	61.20	82.60	77.40	57.69
LingBot-Video-PE	40.60	62.80	46.20	35.60	60.20	65.80	4.60	77.80	70.40	48.66
VPO	0.40	25.20	27.40	64.20	48.80	89.60	6.60	75.20	70.00	33.07
Prompt-A-Video-Cog	0.60	23.00	21.00	58.20	48.40	83.80	14.60	65.20	71.20	29.66
Prompt-A-Video-OS	1.00	19.00	14.40	30.80	22.40	86.80	9.60	37.40	54.40	18.99
RAPO	0.00	1.40	1.80	2.40	1.40	14.40	1.00	3.60	4.20	0.93
I2V-PE										
H3-Context-IR	**34.33**	**73.91**	**75.92**	**79.60**	**87.96**	82.94	**96.99**	**75.25**	61.20	**73.35**
LTX-2.5-PE	2.00	43.67	52.33	73.33	53.33	**91.00**	60.33	65.33	75.00	46.77
LingBot-Video-PE	19.00	50.00	48.33	43.00	38.33	72.00	5.33	67.67	**83.00**	37.47
R2V-PE										
H3-Context-IR	**23.00**	**78.00**	**76.67**	**80.33**	**93.33**	**92.33**	**84.00**	**68.00**	**70.33**	**73.62**
LTX-2.5-PE	5.67	49.67	41.67	62.00	62.00	81.00	51.00	42.33	62.67	38.02

and second in Semantic Fidelity. On the visually conditioned tasks, H3-Context-IR ranks first in all three dimensions on both I2V-PE and R2V-PE. No single method therefore dominates across all tasks and dimensions, revealing distinct capability profiles among current PE.

Finding 1: *Modern prompt enhancement is evolving from fine-grained descriptive expansion toward intent-preserving cinematic planning, enabled by increasingly capable video generators.*

Modern video generators can follow richer, more structured instructions for temporal progression, camera design, narrative development, and audiovisual coordination within a coherent generation plan, expanding the role of PE beyond local detail enrichment. This shift is evident in large-scale, industry-developed PE: WanPE establishes a clear lead in Directorial Creation on T2V-PE, while H3-Context-IR leads by substantial margins on I2V-PE and R2V-PE. Appendix B.3 contrasts descriptive expansion with structured cinematic planning across PE systems. Beyond this emerging industrial trend, Internal Consistency remains uniformly high, varying by less than 9 points within each task, whereas Semantic Fidelity and Directorial Creation each span more than 70 points on T2V-PE. These results indicate that the central challenge lies not merely in expanding prompts or maintaining internal coherence, but in cinematic planning that preserves user intent.

Finding 2: *PE trained to reconstruct real-video captions exhibits broader coverage of cinematic language, while forward-refinement methods excel at semantic preservation or internal coherence.*

The dimension-level results on T2V-PE reveal a clear correspondence between enhancement paradigms and capability profiles. WanPE-397B and WanPE-35B are trained on pairs constructed by annotating real videos with a video captioner and reconstructing compatible user requests from the resulting captions, while retaining the full captions as targets. This supervision exposes them to real-world temporal organization, cinematic language, and audiovisual coordination, consistent

Table 3: Temporal-complexity results.

Structure	Count	SF	IC	DC
Event stages	1	69.60	96.66	78.43
	2	68.79	94.02	76.70
	3+	65.93	92.74	76.45
Requested shots	1	72.48	95.81	79.03
	2	69.90	94.52	75.62
	3+	62.51	88.20	73.51

Table 4: Reference-complexity results.

Reference	Group	SF	IC	DC
Images	1	62.01	96.70	77.09
	2	60.37	95.57	76.56
	3+	54.59	92.31	74.38
Modality	Image-only	58.64	94.66	75.88
	Video-only	67.79	94.46	74.75
	Hybrid	45.62	97.09	55.69

with their leading Directorial Creation scores of 78.97 and 78.15 and their strong performance in storyboard and timing, editing and transitions, and sound design. In contrast, SCMaPR combines policy-conditioned rewriting, atom-level semantic verification, and conditional revision, whereas VPO trains a prompt optimizer through principle-based SFT and multi-feedback preference optimization. These mechanisms help preserve user requirements and constrain unsupported or contradictory elaboration, consistent with their respective leadership in Semantic Fidelity (87.92) and Internal Consistency (95.07), but their lower Directorial Creation scores (49.18 and 33.07) indicate narrower cinematic coverage. These complementary profiles suggest that effective PE should combine caption-derived planning knowledge with explicit semantic and logical constraints.

4.3 Analysis

Building on the preceding findings, we analyze the necessity of prompt enhancement, the relationship between output length and adequacy, and the impact of task complexity.

Original prompts reveal a substantial adequacy gap. We evaluate the original prompts as a no-enhancement baseline. Across all tasks, they achieve high Semantic Fidelity but low Directorial Creation. On T2V-PE, they attain 100.00 in Semantic Fidelity but only 33.56 in Directorial Creation, indicating that preserving user intent alone is insufficient to fully specify the intended video. This observation raises the question of whether greater prompt length necessarily yields higher Directorial Creation. A length-stratified analysis shows no consistent benefit: across tasks, the longest output quartile trails the shortest by 5.19-10.09 points (see Appendix C.2 for details). Effective PE should therefore add structured, task-relevant specifications for temporal planning, cinematography, narrative development, and audiovisual design rather than merely lengthening prompts.

Finding 3: *Temporal and reference complexity remain persistent challenges for prompt enhancers.*

To examine these challenges, we use H3-Context-IR as a fixed enhancer and analyze two forms of task complexity: temporal complexity on T2V-PE and reference complexity on R2V-PE.

Temporal-structure complexity. We analyze T2V-PE to isolate temporal complexity from visual-reference grounding. Specifically, we characterize temporal structure by the numbers of ordered event stages and explicitly requested shots, grouping both into 1, 2, and 3+. An event stage denotes a distinct action, state change, or outcome, while the shot analysis includes only prompts with explicit shot structures. Under these groupings, Table 3 shows that cases with three or more event stages trail those with one stage by 3.67, 3.92, and 1.98 points in SF, IC, and DC, respectively. The corresponding gaps for requested shots are larger at 9.97, 7.61, and 5.52 points. Together, these results show that both event progression and multi-shot organization pose temporal challenges, with the latter placing greater demands on cross-segment coordination.

Reference complexity. On R2V-PE, we characterize reference complexity along two axes: quantity and modality. For quantity, we group image-only cases by 1, 2, and 3+ images. For modality, we compare image-only, video-only, and hybrid inputs. As shown in Table 4, cases with three or more images trail single-image cases by 7.42, 4.39, and 2.71 points in SF, IC, and DC, respectively. The modality comparison reveals a larger gap. Hybrid inputs score only 45.62 in SF and 55.69 in DC, substantially below both single-modality settings, despite achieving the highest IC of 97.09. This divergence reveals a reference-grounding bottleneck. Enhanced prompts can remain internally coherent while failing to faithfully and adequately integrate heterogeneous references.

Qualitative failure analysis. As shown in Figure 4, evaluated PE exhibit representative failures across all three dimensions. The corresponding videos generated by Wan3.0 visibly reflect these errors, illustrating the interpretability and downstream relevance of our prompt-level evaluation.



Figure 4: **Qualitative failure cases of evaluated prompt enhancers, visualized with Wan3.0 video model.** Examples include (a) visual-style deviation, (b) incorrect subject–action binding, (c) language and speaker attribution errors, (d) narrative truncation caused by inadequate temporal planning, and (e) incorrect grounding across multiple references.

Table 5: Spearman correlations for human alignment and downstream utility across three tasks.

	T2V-PE			I2V-PE			R2V-PE		
Evaluation	SF	IC	DC	SF	IC	DC	SF	IC	DC
Enhanced-prompt alignment									
Direct VLM Scoring	0.618	0.464	0.527	1.000	0.600	0.400	1.000	0.500	0.500
WanPEBench Evaluation	0.955	0.818	0.909	1.000	1.000	1.000	1.000	1.000	1.000
Downstream-video alignment									
WanPEBench Evaluation (Wan3.0)	0.845	0.827	0.945	1.000	1.000	1.000	1.000	1.000	1.000
WanPEBench Evaluation (MiniMax-H3)	0.809	0.864	0.936	1.000	1.000	1.000	1.000	1.000	1.000

4.4 Human Alignment

Finding 4: *WanPEBench evaluation scores reliably reflect expert judgments of both enhanced prompts and downstream videos across generators.*

To validate the reliability of our evaluation framework, we measure Spearman correlations (ρ_s) with expert rankings at the enhanced-prompt and downstream-video levels. The resulting correlations are reported in Table 5. At the enhanced-prompt level, our framework aligns more closely with expert rankings than direct VLM scoring, with dimension-averaged gains of 0.333-0.358 across tasks. This improvement supports the effectiveness of rubric-based decomposition. At the downstream-video level, we compare the same evaluation scores with expert rankings of downstream videos generated by Wan3.0 and MiniMax-H3. The resulting correlations of 0.809-1.000 indicate that direct prompt evaluation reliably reflects downstream utility across generators. The study covers all PE systems in Table 2, with the original prompts as a baseline. Annotation details are provided in Appendix D.

5 Conclusion

We present WanPEBench, the first unified benchmark for direct evaluation of prompt enhancers across T2V-PE, I2V-PE, and R2V-PE. It comprises 1,100 expert-verified cases and 1,005 visual assets spanning 35 fine-grained tasks encompassing diverse temporal, cinematic, audiovisual, and multi-reference requirements. Its evidence-grounded framework combines modality-aware evidence with 24 criterion-specific rubrics, enabling consistent, generator-independent assessment without costly video rendering. Our evaluation reveals an emerging shift from descriptive expansion toward intent-preserving cinematic planning. At the method level, caption-reconstruction and forward-refinement methods exhibit complementary strengths, suggesting that future PE systems should combine cinematic knowledge with semantic and logical safeguards. Meanwhile, performance degradation under temporal and multi-reference complexity highlights persistent challenges in temporal planning and cross-asset grounding. Strong alignment with expert judgments at both the prompt and downstream-video levels confirms the practical relevance of direct PE evaluation. By unifying task coverage, evidence-grounded assessment, and criterion-level analysis, WanPEBench provides a scalable foundation for evaluating, comparing, and advancing video prompt enhancers.

AI USE STATEMENT

In accordance with ICLR 2027 policy, we report the use of AI models in this work. The AI models used for dataset construction and evaluation are described in the main paper. Beyond these methodological uses, AI assistance during manuscript preparation was strictly limited to enhancing linguistic quality, including improving sentence structure and ensuring grammatical correctness. The entirety of the scientific contributions in this paper is attributable solely to the authors. The final manuscript has been critically reviewed by all authors, who collectively assume full responsibility for its accuracy and integrity.

REFERENCES

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

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, 2025. URL https://arxiv.org/abs/2503.20491.

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 2025a.

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. *arXiv preprint arXiv:2510.20206*, 2025b.

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.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, and Yongxin Ni. Video-Bench: Human-Aligned Video Generation Benchmark . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18858–18868, Los Alamitos, CA, USA, June 2025. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.01757. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.01757.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023. URL https://arxiv.org/abs/2311.17982.

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, 2024. URL https://arxiv.org/abs/2412.15156.

Yizhuo Jia, Jingyun Hua, and Yuanxing Zhang. Cape-t2v: Captioner-anchored prompt enhancement toward two-sided conditioning alignment in text-to-video generation, 2026. URL https://arxiv.org/abs/2608.03046.

Sen Liang, Cong Wang, Zhentao Yu, Fengbin Guan, Zhengguang Zhou, Teng Hu, Youliang Zhang, Yuan Zhou, Xin Li, Qinglin Lu, and Zhibo Chen. Goku: A million-scale universal dataset and benchmark for instruction-based video editing, 2026. URL https://arxiv.org/abs/2606.30599.

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

Kaiqi Liu, Haoxuan Zeng, Jingqi Liu, Jiacong Fang, Ziqi Cai, Yunyao Mao, Henglin Liu, Yu Sheng, Shuchen Weng, and Boxin Shi. Streamav-bench: A comprehensive benchmark for streaming audio-video generation, 2026. URL https://arxiv.org/abs/2608.26336.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models, 2024. URL https://arxiv.org/abs/2310.11440.

Yexin Liu, Manyuan Zhang, Yueze Wang, Hongyu Li, Dian Zheng, Weiming Zhang, Changsheng Lu, Xunliang Cai, Yan Feng, Peng Pei, and Harry Yang. Opensubject: Leveraging video-derived identity and diversity priors for subject-driven image generation and manipulation, 2025. URL https://arxiv.org/abs/2512.08294.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö. Arık. Vista: A test-time self-improving video generation agent, 2025. URL https://arxiv.org/abs/2510.15831.

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.

Chaojie Mao, Chen-Wei Xie, Chongyang Zhong, Haoyou Deng, Jiaxing Zhao, Jie Xiao, Jinbo Xing, Jingfeng Zhang, Jingren Zhou, Jingyi Zhang, Jun Dan, Kai Zhu, Kang Zhao, Keyu Yan, Minghui Chen, Pandeng Li, Shuangle Chen, Tong Shen, Yu Liu, Yue Jiang, Yulin Pan, Yuxiang Tuo, Zeyinzi Jiang, Zhen Han, Ang Wang, Bang Zhang, Baole Ai, Bin Wen, Boang Feng, Feiwu Yu, Gang Wang, Haiming Zhao, He Kang, Jianjing Xiang, Jianyuan Zeng, Jinkai Wang, Junjie Zhou, Ke Sun, Linqian Wu, Pei Gong, Pingyu Wu, Ruiwen Wu, Tongtong Su, Wenmeng Zhou, Wenting Shen, Wenyuan Yu, Xianjun Xu, Xiaoming Huang, Xiejie Shen, Xin Xu, Yan Kou, Yangyu Lv, Yifan Zhai, Yitong Huang, Yun Zheng, Yuntao Hong, Zhe Zhang, and Zhicheng Zhang. Wan-image: Pushing the boundaries of generative visual intelligence, 2026. URL https://arxiv.org/abs/2604.19858.

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

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation, 2025. URL https://arxiv.org/abs/2407.02371.

OpenAI. Gpt-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026a.

OpenAI. Gpt-image-2. https://openai.com/index/introducing-chatgpt-images-2-0/, 2026b.

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

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget,

Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

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

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 2025a. doi: 10.24963/ijcai.2025/1133. URL https://doi.org/10.24963/ijcai.2025/1133. AI, Arts & Creativity.

Linqing Wang, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Jiale Tao, Qixun Wang, Ruihuang Li, Comi Chen, Xin Li, Mingrui Wu, Xinchi Deng, Shuyang Gu, Chunyu

Wang, and Qinglin Lu. Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting, 2025b. URL https://arxiv.org/abs/2509.04545.

Jianhui Wei, Xiaotian Zhang, Yichen Li, Yuan Wang, Yan Zhang, Ziyi Chen, Zhihang Tang, Wei Xu, and Zuozhu Liu. Univbench: Towards unified evaluation for video foundation models, 2026a. URL https://arxiv.org/abs/2602.21835.

Yujie Wei, Yujin Han, Zhekai Chen, Yongming Li, Kaixun Jiang, Zhihang Liu, Quanhao Li, Zhiwu Qing, Xiang Wang, Zhen Xing, Ruihang Chu, Lingyi Hong, Yefei He, Junjie Zhou, Junqiu Yu, Yang Shi, Difan Zou, Kai Zhu, Shiwei Zhang, Yingya Zhang, Yu Liu, Xihui Liu, and Hongming Shan. Msavbench: Towards comprehensive and reliable evaluation of multi-shot audio-video generation, 2026b. URL https://arxiv.org/abs/2605.20183.

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, 2026. URL https://arxiv.org/abs/2604.05489.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation, 2025. URL https://arxiv.org/abs/2505.20292.

Yubo Zhu, Yawen Shao, Ziyun Dai, Zixun Fang, Kai Zhu, Siyang Sun, Haolan Xue, Chuxin Wang, Tingyu Weng, Jingming Luo, Chen Shi, Lianghua Huang, Yufeng Ai, Yuzheng Wang, Wenyuan Zhang, Yu Shang, Yuxiang Bao, Zoubin Bi, Jie Xiao, Jinbo Xing, Jiaxing Zhao, Chongyang Zhong, Hengjian Chen, Chenwei Xie, Akide Liu, Zhehan Kan, Yu Liu, Wei Zhai, Sheng Zhong, and Wei Tong. Wanpe: Towards cinematic prompt enhancement for modern text-to-video generation, 2026. URL https://arxiv.org/abs/2609.30221.

APPENDIX OVERVIEW

This appendix supplements the main paper with data, evaluation, experimental, and human annotation details, followed by ethical considerations and limitations. It is organized as follows.

- **Appendix A: More Data Details.** Data design, construction, and detailed analysis.
- **Appendix B: More Evaluation Details.** Evaluation dimensions and criteria, structured output formats, and representative case.
- **Appendix C: Additional Experimental Details.** Implementation details and length-stratified analysis.
- **Appendix D: Human Expert Annotation.** Pairwise annotation protocol, quality control, and annotation interfaces.
- **Appendix E: Ethics, Privacy, and Licensing.** Data screening, asset licensing, and release policy.
- **Appendix F: Limitations.** Benchmark scale and future expansion.

A MORE DATA DETAILS

A.1 DATA DESIGN DETAILS

Prompt enhancement settings. The benchmark unifies three settings with progressively richer conditioning. T2V-PE contains 500 text-only cases. I2V-PE contains 300 cases, each paired with one first-frame image. R2V-PE contains 300 cases conditioned on one or more images or videos whose roles may include subject identity, object, scene, style, motion, cinematography, transition pattern, or audiovisual rhythm. Together, these settings cover prompt enhancement from text, temporal continuation from a first frame, and integration of multiple references.

Tasks. The benchmark defines a two-level task taxonomy with 9 high-level capability categories and 35 fine-grained task types. Each task type belongs to one parent category. Table 6 lists the taxonomy and the number of task types in each category.

Table 6: **Hierarchical task taxonomy of WanPEBench.** The nine high-level categories contain 35 fine-grained task types.

High-level category	#	Fine-grained task types
Action	4	Daily Activity, Combat, Sports, Dance
Animation	2	2D Animation, 3D Animation
Visual Effects	5	VFX Character, VFX Object, Transformation, Light/Weather Effects, Sci-Fi Creature
General	5	Physical Laws, On-Screen Text, Portrait, Nature, Intellectual Property (IP)
Narrative	3	Modern Story, Historical Story, Fantasy Story
Cinematography	3	Camera Motion, Time-Lapse, Lighting and Color
Speech	5	Dialogue, Monologue, Narration, Explanation, Emotional Vocalization
Music	6	Singing, Song and Dance, Music Video, Opera, Background Music, Instrumental Performance
Sound Effects	2	Animal Sound, Environmental Sound

Scenario composition dimensions. Each scenario is composed along four additional axes. The five subject categories are Human, Animal, Object, Fantasy Being, and No Salient Subject. The eight scene domains are Domestic, Urban, Workplace, Performance, Sports, Transportation, Nature, and Otherworldly. The seven visual style classes are Photorealistic, Anime, Graphic Illustration, 3D CGI, Painting, Digital Stylized, and Mixed/Unspecified. The eight explicit audio domains are Speech and Vocalization, Singing, Music, Human Activity Sounds, Foley and Object Sounds, Natural and Animal Sounds, Environmental Ambience, and Mechanical and Vehicle Sounds.

Linguistic, temporal, and cinematic dimensions. The multilingual portion spans English, Chinese, Japanese, Korean, Spanish, French, Arabic, Russian, Bengali, German, Portuguese, and Hindi. Prompts vary from concise concepts to detailed scripts and may describe a single continuous shot

or as many as 13 explicitly structured shots. Cinematic coverage includes near, medium, and long shot scale families. It also covers level, high, low, overhead or aerial, and subjective or special viewpoints. Lens families include wide, standard, telephoto, and specialty lenses. Camera motion includes translational, rotational, and compound movement. Composition, focus, transition, lighting, palette, and color grading provide additional controls that may overlap.

A.2 Data Construction Details

Figure 2 summarizes the four-stage construction pipeline. This section describes the design choices and quality checks used at each stage.

Stage 1: Taxonomy Design and Composition

Domain experts define the nine capability categories and 35 tasks in Table 6. The taxonomy covers visual, narrative, cinematic, speech, music, and sound capabilities. It therefore tests both frame content and the development of events, camera behavior, and audio over time. The experts also define the subject, scene, style, and audio vocabularies described in Section A.1. Each candidate scenario is represented as a $(\text{category}, \text{subject}, \text{scene}, \text{style}, \text{audio})$ quintuple. It is then paired with a task, a prompt enhancement setting, and a target duration. Incompatible combinations are discarded rather than forced. Coverage is monitored across the dataset to maintain diversity without producing incoherent or infeasible requests.

Stage 2: Prompt Synthesis and Diversification

Gemini 3.1 Pro synthesizes a candidate user request from the sampled specification. It considers the task and conditioning modality, the scenario composition, and the content that can reasonably unfold within the target duration. Instruction specificity ranges from short concepts to descriptive requests and detailed scripts. Temporal structure also varies. Some requests leave the shot structure implicit, while others specify ordered events, timestamps, or multiple shots. This process varies how much information and temporal planning are provided to the prompt enhancer.

Stage 3: Reference Asset Acquisition

Cases with reference inputs use two acquisition routes. Public data selection draws candidate images and videos from existing datasets. Images are screened for visual quality, content safety, resolution, and redundancy. Videos are screened for visual quality and motion prominence, segmented into semantically complete clips, and represented by keyframes when a static reference is needed.

Generation with models expands the public pool using Wan-Image (Mao et al., 2026), Seedream (Seedream et al., 2025), and GPT Image 2 (OpenAI, 2026b). In the image guided route, a primary image is used to create variations in viewpoint, pose, or expression while preserving identity. In the text guided route, Gemini 3.1 Pro converts the sampled scenario or a relevant shot description into a concise image generation instruction. Together, the two routes support single image, multiple image, video only, and mixed image and video conditioning. Each retained reference is assigned an explicit role such as subject, scene, style, motion, or cinematography.

Stage 4: Automated Verification and Expert Curation

Quality control combines automated screening with expert review. Gemini 3.1 Pro derives a checklist from each request and its target duration. It checks task alignment, logical coherence, and duration feasibility. For cases with reference inputs, it jointly inspects the instruction and candidate assets to verify their assigned roles, alignment with the script, reference coverage, and consistency across assets.

Ten domain experts review the verified pool. They refine awkward language, remove ambiguous or infeasible instructions, reject uninformative or mismatched references, and eliminate repetitive or redundant cases. They also monitor task balance, difficulty, and novelty across cases. The final benchmark contains 1,100 cases verified by experts and 1,005 visual assets.

A.3 Detailed Data Analysis

Task coverage. WanPEBench covers 9 high level capability categories: Action (25.3%), Visual Effects (16.0%), Narrative (13.4%), Animation (10.0%), Speech (9.7%), General (9.1%), Music (7.6%), Cinematography (6.8%), and Sound Effects (2.1%).

Subject, scene, and style diversity. Subjects span 5 categories: Human (50.4%), Object (16.6%), Fantasy Being (14.4%), Animal (10.1%), and No Salient Subject (8.5%). Scenes span 8 categories: Otherworldly (19.8%), Domestic (14.5%), Urban (14.1%), Workplace (14.1%), Nature (13.2%), Performance (11.9%), Sports (7.5%), and Transportation (5.0%). T2V-PE visual styles span 7 categories: Photorealistic (61.2%), Digital Stylized (10.4%), 3D CGI (8.2%), Mixed/Unspecified (8.2%), Graphic Illustration (5.0%), Painting (3.8%), and Anime (3.2%).

Audio and language diversity. Prompts cover 12 languages: English (50.0%), Chinese (40.0%), Japanese (1.0%), Korean (1.0%), Spanish (1.0%), French (1.0%), Arabic (1.0%), Russian (1.0%), Bengali (1.0%), German (1.0%), Portuguese (1.0%), and Hindi (1.0%). Audio content spans 8 categories: Speech and Vocalization (38.3%), Music (19.9%), Foley and Object Sounds (14.6%), Singing (12.6%), Natural and Animal Sounds (7.7%), Mechanical and Vehicle Sounds (3.1%), Human Activity Sounds (2.7%), and Environmental Ambience (1.1%).

Prompt length. Prompt length spans 7 intervals: at most 50 tokens (14.1%), 51 to 100 tokens (13.8%), 101 to 200 tokens (17.7%), 201 to 400 tokens (20.1%), 401 to 800 tokens (16.4%), 801 to 1,600 tokens (15.9%), and more than 1,600 tokens (2.0%). Table 7 reports detailed token length statistics for each task. Target durations range from 2 to 30 seconds. There are 147 cases from 2 to 5 seconds, 321 from 6 to 10 seconds, 614 from 11 to 15 seconds, 11 from 16 to 20 seconds, and 7 from 21 to 30 seconds.

Table 7: **Input prompt token length statistics.** Quartiles are computed independently within each setting.

Setting	Min.	Q1	Median	Mean	Q3	Max.
T2V-PE	6	85.3	237.5	425.8	597.3	3,431
I2V-PE	8	91.8	221.5	448.9	642.8	4,120
R2V-PE	13	89.8	210.5	363.8	549.0	2,412
All	6	88.8	224.0	415.2	596.8	4,120

Cinematic language. WanPEBench reports 3 major shot scale families: near (49.5%), long (30.7%), and medium (19.8%). It covers 5 major camera angles: eye level (38.0%), high (23.2%), low (20.2%), subjective (13.7%), and others (4.9%). It also covers 4 lens categories: wide angle (38.1%), telephoto or zoom (27.6%), specialty (25.4%), and standard (9.0%). Camera movement spans 3 types: pure translation (50.7%), compound (44.2%), and pure rotation (5.0%).

Reference diversity and complexity. R2V-PE reference configurations comprise 3 types: image only (80.0%), video only (9.3%), and mixed image and video (10.7%). Cases with multiple images span 6 reference counts: two (39.2%), three (21.7%), four (19.3%), five (16.9%), six (1.8%), and seven (1.2%). Multi-image reference structures comprise 5 types: multi view (5.4%), multiple subjects (15.1%), subject with object or scene (56.0%), sequential frames (21.7%), and others (1.8%). Reference asset styles comprise 3 groups: realistic (54.6%), nonrealistic (24.8%), and mixed or unspecified (20.6%). In absolute terms, I2V-PE contributes 300 first frame images. R2V-PE contributes 644 images and 61 videos. Its image conditioned cases include 106 with one image, 65 with two, 36 with three, 32 with four, 28 with five, 3 with six, and 2 with seven. Its video conditioned cases include 59 with one video and 1 with two videos.

B More Evaluation Details

B.1 Evaluation Dimensions and Criteria

WanPEBench evaluates enhanced prompts using 24 criteria grouped into Semantic Fidelity (*SF*), Internal Consistency (*IC*), and Directorial Creation (*DC*). For each applicable criterion, the evaluator

determines whether the enhanced prompt contains no issue, a minor issue, or a major issue, and cites evidence for that decision. A criterion is counted as passed only when no issue is identified. Criterion-level pass rate is therefore the percentage of applicable cases that pass the criterion, while each dimension-level score aggregates the number and severity of its constituent issues. Criteria tied to optional source content, such as dialogue or on-screen text, are applied only when that content is present or explicitly requested.

B.1.1 SEMANTIC FIDELITY METRICS

Semantic Fidelity evaluates whether the enhanced prompt preserves the user instruction and instruction-relevant visual evidence. Added details must not omit, weaken, substitute, or contradict source requirements.

(1) Genre and Style Fidelity (GSF). Measures preservation of the requested genre, medium, aesthetic, and visual style. Replacing the medium, weakening a defining genre convention, or adding an incompatible reference style constitutes an issue.

(2) Subject and Attribute Preservation (SAP). Measures preservation of subject identity, count, appearance, attributes, objects, and their bindings. I2V uses the first frame as the initial-state anchor. R2V follows each reference's assigned role.

(3) Action and Narrative Fidelity (ANF). Measures preservation of required actions, interactions, event order, outcomes, and narrative intent. Omissions, substitutions, reversals, or added events that alter the requested event chain are treated as issues.

(4) Dialogue and Lyrics Integrity (DLI). Measures fidelity to dialogue, narration, lyrics, language, delivery, and speaker attribution. Exact quotations must not be paraphrased, deleted, translated, or converted into another form unless explicitly permitted.

(5) Audio Requirement Preservation (ARP). Measures preservation of speech, vocalization, music, sound effects, ambience, and silence constraints. Missing, substituted, or added sounds that violate negative requirements such as "no music" are penalized.

(6) Camera Constraint Preservation (CCP). Measures preservation of shot count, scale, angle, lens, focus, movement, and single-take or static-camera constraints. Added camera design is allowed only where it does not override explicit cinematic intent.

(7) Lighting and Color Fidelity (LCF). Measures preservation of time of day, illumination, light quality, brightness, palette, saturation, contrast, and grading. Compatible elaboration is allowed, but it must not change a defining lighting or color condition.

(8) Scene and Spatial Preservation (SSP). Measures preservation of setting, background elements, layout, relative positions, and movement directions. Changing the location, removing required objects, or reversing spatial relations is a violation.

(9) On-Screen Text Integrity (OTI). Measures fidelity to visible titles, signs, subtitles, and interface text, including exact wording, spelling, language, placement, and timing. Visible text cannot be replaced by semantically similar spoken content.

B.1.2 INTERNAL CONSISTENCY METRICS

Internal Consistency evaluates whether the complete enhanced prompt is logically, spatially, temporally, physically, and audiovisually coherent.

(10) Subject and Attribute Continuity (SAC). Measures stability of subject identity, count, appearance, clothing, props, and persistent attributes across shots. Unexplained renaming, duplication, disappearance, or attribute drift constitutes an issue.

(11) Action and Narrative Logic (ANL). Measures whether actions form a coherent causal or procedural progression. The evaluator checks prerequisites, bindings between subjects and actions, reactions, resolution, and incompatible or dangling events.

(12) Temporal and State Continuity (TSC). Measures chronology, duration allocation, event order, and state transitions. Overlapping or infeasible timestamps, unexplained resets, and later shots showing states that should no longer exist are penalized.

(13) Spatial Consistency (SPC). Measures consistency of location, screen direction, relative position, orientation, trajectory, and entrances or exits. Unexplained position swaps or movement paths that cannot connect adjacent states constitute issues.

(14) Physical Logic (PL). Measures compliance with real-world physics or established fictional rules, including gravity, contact, object permanence, bodily feasibility, and causal relations. Deliberate fantasy is judged by its own world logic.

(15) Audiovisual Synchronization (AVS). Measures temporal and semantic alignment between visible events and dialogue, lyrics, music, effects, or ambience. Wrong speakers, premature effects, and misaligned lip-sync or beat-sync actions are penalized.

B.1.3 Directorial Creation Metrics

Even casually captured videos involve temporal progression, framing, camera behavior, lighting, performance, visual style, and sound. A prompt may preserve the source and remain logically coherent while leaving these directorial elements underspecified. To assess whether a prompt enhancer develops them into an actionable production plan, we introduce Directorial Creation, covering storyboarding, cinematography, editing, lighting, narrative, sound, performance, and visual style.

(16) Storyboard and Timing Completeness (STC). Measures whether the content is organized into a feasible temporal plan with adequate event or shot boundaries, ordering, and durations.

(17) Camera Design Specificity (CDS). Measures actionable specs of framing, scale, angle, viewpoint, lens, focus, and movement.

(18) Editing and Transition Design (ETD). Measures whether cuts and transitions support temporal, spatial, narrative, or rhythmic progression. Disconnected shots or repeated transitions without clear continuity or motivation are considered inadequate.

(19) Lighting and Color Specification (LCS). Measures actionable specification of light source, direction, quality, contrast, palette, saturation, and grading.

(20) Narrative Development and Engagement (NDE). Measures sufficient progression, variation, escalation, or resolution for the requested content form. Static restatement and unmotivated repetition are inadequate.

(21) Core-Intent Focus (CIF). Measures whether additions support the central request. Repetitive padding, template language, unnecessary subplots, and decorative details that distract from the intended result reduce adequacy.

(22) Sound Design Integration (SDI). Measures whether speech, music, effects, and ambience are integrated through source, character, intensity, progression, and visual timing. It rewards content-appropriate design rather than requiring every audio type.

(23) Subject and Performance Specificity (SPS). Measures generation-relevant subject and performance detail, including appearance, pose, expression, gesture, movement, interaction, gaze, and vocal delivery.

(24) Visual Style Specification (VSS). Measures actionable specification of medium, texture, rendering language, production design, and visual treatment.

B.2 Structured Output Formats

B.2.1 Asset-Fact Extraction Output

We present R2V as the representative example because it contains the most complete multimodal evidence structure. Qwen3.5-Omni-Plus (Team, 2026) returns the following fixed *AssetFacts* structure, with each fact traceable to its source asset.

R2V AssetFacts Output Format

```
{
  "asset_inventory": {
```

```
    "total_count": 0,
    "image_count": 0,
    "video_count": 0,
    "items": [
      {
        "asset_id": "image_1|video_1",
        "asset_type": "image|video",
        "input_index": 1,
        "modality_index": 1
      }
    ]
  },
  "assets": [
    {
      "asset_id": "image_1|video_1",
      "asset_type": "image|video",
      "duration_seconds": "xxx",
      "media_and_style": {
        "media_type": "live_action|2d_animation|3d_cg|stop_motion|illustration|game_render
    |mixed",
        "realism": "realistic|stylized_realistic|non_realistic",
        "primary_style": "xxx [dominant visual style]",
        "secondary_styles": ["xxx [additional visual style]"],
        "visible_style_evidence": ["xxx [visible evidence for the style]"],
        "certainty": "certain|uncertain",
        "uncertainty_reason": "xxx [reason when uncertain]"
      },
      "overall_content_summary": "xxx [brief summary of the asset]",
      "subjects": [
        {
          "id": "image_1_subject_1|video_1_subject_1",
          "role": "primary|secondary|background_group",
          "category": "xxx [person, animal, object, or group]",
          "identity": "xxx [visible identity description]",
          "count_exact": 1,
          "count_estimate": "xxx",
          "appearance": ["xxx [visible appearance detail]"],
          "pose_and_current_state": ["xxx [visible pose or current state]"],
          "expression": "xxx [visible expression if identifiable]",
          "held_or_worn_items": ["xxx [held or worn item]"],
          "frame_position": "xxx [location within the frame]",
          "occlusion": "none|partial|heavy",
          "certainty": "certain|uncertain",
          "uncertainty_reason": "xxx [reason when uncertain]"
        }
      ],
      "scene_and_spatial_relations": {
        "scene_type": "xxx [visible scene description]",
        "indoor_outdoor": "indoor|outdoor|mixed",
        "foreground_elements": ["xxx [foreground element]"],
        "midground_elements": ["xxx [midground element]"],
        "background_elements": ["xxx [background element]"],
        "spatial_relations": [
          {
            "subject": "xxx [subject ID or entity]",
            "relation": "xxx [visible spatial relation]",
            "object": "xxx [related object or entity]",
            "certainty": "certain|uncertain"
          }
        ],
        "cropped_or_offscreen_information": ["xxx [cropped or offscreen cue]"]
      },
      "camera_and_composition": {
        "shot_scale": "extreme_close_up|close_up|medium_close_up|medium_shot|
    medium_full_shot|full_shot|long_shot|extreme_long_shot|varied",
        "camera_angle": "xxx [camera angle]",
        "camera_height": "xxx [camera height]",
        "view_characteristics": "wide_angle_feel|standard_view|telephoto_compression|
    fisheye|varied",
        "composition": ["xxx [composition characteristic]"],
        "visual_focus": "xxx [primary visual focus]",
        "depth_of_field": "shallow|moderate|deep|varied",
        "camera_motion": ["xxx [observed camera motion]"],
        "certainty": "certain|uncertain",
        "uncertainty_reason": "xxx [reason when uncertain]"
      },
      "lighting_and_color": {
        "light_sources": ["xxx [visible light source]"],
        "light_direction_and_effect": ["xxx [light direction and visible effect]"],
```

```
      "time_of_day_cue": "xxx [visible time-of-day cue if present]",
      "color_temperature": "warm|neutral|cool|mixed",
      "dominant_colors": ["xxx [dominant color]"],
      "accent_colors": ["xxx [accent color]"],
      "certainty": "certain|uncertain",
      "uncertainty_reason": "xxx [reason when uncertain]"
    },
    "onscreen_text": {
      "present": false,
      "items": [
        {
          "text": "xxx [visible text transcription]",
          "type": "subtitle|title|logo|sign|interface|other",
          "language": "xxx [text language]",
          "frame_position": "xxx [text location in the frame]",
          "time_range": {
            "start_seconds": "xxx",
            "end_seconds": "xxx"
          },
          "certainty": "certain|uncertain",
          "uncertain_fragments": ["xxx [uncertain text fragment]"]
        }
      ]
    },
    "segments": [
      {
        "segment_id": "video_1_segment_1",
        "shot_index": 1,
        "is_new_shot": true,
        "start_seconds": "xxx",
        "end_seconds": "xxx",
        "transition_in": "xxx [incoming transition if present]",
        "scene": "xxx [scene during the segment]",
        "subjects_present": ["xxx [visible subject ID]"],
        "visible_actions_and_state_changes": ["xxx [visible action or state change]"],
        "subject_interactions": ["xxx [visible subject interaction]"],
        "object_events": ["xxx [visible object event]"],
        "camera": {
          "shot_scale": "xxx [segment shot scale]",
          "angle": "xxx [segment camera angle]",
          "camera_height": "xxx [segment camera height]",
          "view_characteristics": "xxx [segment lens characteristic]",
          "composition": ["xxx [segment composition]"],
          "focus": "xxx [segment visual focus]",
          "depth_of_field": "xxx [segment depth of field]",
          "movement": {
            "type": "static|push_in|pull_out|pan|tilt|track|crane|orbit|zoom|handheld|
compound",
            "direction": "xxx [movement direction]",
            "speed": "xxx [movement speed]",
            "amplitude": "xxx [movement amplitude]",
            "stability": "xxx [movement stability]",
            "tracking_subject": "xxx [tracked subject ID if applicable]",
            "start_frame": "xxx [movement start state]",
            "end_frame": "xxx [movement end state]"
          }
        },
        "shot_change_from_previous": {
          "shot_scale_change": "xxx [change in shot scale]",
          "angle_change": "xxx [change in camera angle]",
          "camera_position_change": "xxx [change in camera position]",
          "composition_change": "xxx [change in composition]",
          "focus_change": "xxx [change in visual focus]",
          "movement_change": "xxx [change in camera movement]",
          "scene_change": "xxx [change in scene]",
          "subject_spatial_change": "xxx [change in subject position]",
          "continuity_basis": ["xxx [evidence supporting continuity]"]
        },
        "visual_effects": ["xxx [visible effect]"],
        "onscreen_text_refs": ["xxx [related text item ID]"],
        "audio_event_refs": ["xxx [related audio event ID]"],
        "certainty": "certain|uncertain",
        "uncertainty_reason": "xxx [reason when uncertain]"
      }
    ],
    "subject_tracks": [
      {
        "subject_id": "video_1_subject_1",
        "first_segment": "video_1_segment_1",
```

```
          "last_segment": "video_1_segment_1",
          "stable_identity_anchors": ["xxx [stable identity cue]"],
          "actions_in_order": ["xxx [ordered subject action]"],
          "state_changes_in_order": ["xxx [ordered subject state change]"],
          "certainty": "certain|uncertain",
          "uncertainty_reason": "xxx [reason when uncertain]"
        }
      ],
      "audio": {
        "present": false,
        "speech": [
          {
            "id": "video_1_audio_1",
            "time_range": {
              "start_seconds": "xxx",
              "end_seconds": "xxx"
            },
            "speaker_subject_id": "xxx [speaker subject ID if identifiable]",
            "source_type": "visible_speaker|offscreen_speaker|voiceover|unknown",
            "language": "xxx [spoken language]",
            "transcript": "xxx [verbatim speech transcript]",
            "vocal_characteristics": ["xxx [audible vocal characteristic]"],
            "certainty": "certain|uncertain",
            "uncertainty_reason": "xxx [reason when uncertain]"
          }
        ],
        "music": [
          {
            "id": "video_1_music_1",
            "time_range": {
              "start_seconds": "xxx",
              "end_seconds": "xxx"
            },
            "description": ["xxx [audible music characteristic]"],
            "certainty": "certain|uncertain",
            "uncertainty_reason": "xxx [reason when uncertain]"
          }
        ],
        "ambient_sounds": ["xxx [audible ambient sound]"],
        "sound_effects": ["xxx [audible sound effect]"],
        "audio_visual_relations": ["xxx [audio-visual synchronization relation]"]
      },
      "uncertainties": ["xxx [asset-level uncertainty]"]
    }
  ],
  "cross_asset_relations": [
    {
      "source": {
        "asset_id": "image_1",
        "entity_id": "image_1_subject_1"
      },
      "target": {
        "asset_id": "video_1",
        "entity_id": "video_1_subject_1"
      },
      "relation": "same_entity|possibly_same_entity|same_scene|possibly_same_scene|
    same_object|possibly_same_object|style_similarity|other",
      "evidence": ["xxx [evidence for the cross-asset relation]"],
      "certainty": "certain|uncertain",
      "uncertainty_reason": "xxx [reason when uncertain]"
    }
  ],
  "asset_anchor_summary": ["xxx [instruction-relevant asset anchor]"],
  "uncertainties": ["xxx [global uncertainty]"]
}
```

B.2.2 THREE-DIMENSION EVALUATION OUTPUTS

The evaluator uses separate fixed outputs for Semantic Fidelity, Internal Consistency, and Directorial Creation.

Semantic Fidelity Output Format

```
{
  "overall_verdict": "good|acceptable|bad",
  "semantic_fidelity_score": "xxx [score from 0 to 100]",
```

```
  "primary_issue": "genre_style|subject|action_narrative|dialogue|sound|camera|
     lighting_color|scene|onscreen_text|none",
  "summary": "",
  "dimensions": {
    "genre_style":      {"status": "consistent|lossy|altered|not_present", "severity": "
     none|minor|major", "evidence_before": "", "evidence_after": "", "note": ""},
    "subject":          {"status": "consistent|lossy|altered|not_present", "severity": "
     none|minor|major", "evidence_before": "", "evidence_after": "", "note": ""},
    "action_narrative": {"status": "consistent|lossy|altered|not_present", "severity": "
     none|minor|major", "evidence_before": "", "evidence_after": "", "note": ""},
    "dialogue":         {"status": "consistent|lossy|altered|not_present", "severity": "
     none|minor|major", "evidence_before": "", "evidence_after": "", "note": ""},
    "sound":            {"status": "consistent|lossy|altered|not_present", "severity": "
     none|minor|major", "evidence_before": "", "evidence_after": "", "note": ""},
    "camera":           {"status": "consistent|lossy|altered|not_present", "severity": "
     none|minor|major", "evidence_before": "", "evidence_after": "", "note": ""},
    "lighting_color":   {"status": "consistent|lossy|altered|not_present", "severity": "
     none|minor|major", "evidence_before": "", "evidence_after": "", "note": ""},
    "scene":            {"status": "consistent|lossy|altered|not_present", "severity": "
     none|minor|major", "evidence_before": "", "evidence_after": "", "note": ""},
    "onscreen_text":    {"status": "consistent|lossy|altered|not_present", "severity": "
     none|minor|major", "evidence_before": "", "evidence_after": "", "note": ""}
  }
}
```

Internal Consistency Output Format

```
{
  "overall_verdict": "good|acceptable|bad",
  "internal_consistency_score": "xxx [score from 0 to 100]",
  "primary_issue": "subject_continuity|action_narrative_logic|temporal_state_continuity|
     spatial_consistency|physical_plausibility|audio_visual_sync|none",
  "summary": "",
  "dimensions": {
    "subject_continuity":         {"status": "consistent|ambiguous|inconsistent|
     not_present", "severity": "none|minor|major", "evidence": "", "note": ""},
    "action_narrative_logic":     {"status": "consistent|ambiguous|inconsistent|
     not_present", "severity": "none|minor|major", "evidence": "", "note": ""},
    "temporal_state_continuity":  {"status": "consistent|ambiguous|inconsistent|
     not_present", "severity": "none|minor|major", "evidence": "", "note": ""},
    "spatial_consistency":        {"status": "consistent|ambiguous|inconsistent|
     not_present", "severity": "none|minor|major", "evidence": "", "note": ""},
    "physical_plausibility":      {"status": "consistent|ambiguous|inconsistent|
     not_present", "severity": "none|minor|major", "evidence": "", "note": ""},
    "audio_visual_sync":          {"status": "consistent|ambiguous|inconsistent|
     not_present", "severity": "none|minor|major", "evidence": "", "note": ""}
  }
}
```

Directorial Creation Output Format

```
{
  "overall_verdict": "good|acceptable|bad",
  "directorial_creation_score": "xxx [score from 0 to 100]",
  "major_criterion_count": 0,
  "minor_criterion_count": 0,
  "primary_strength": "storyboard_timing|camera_design|editing_transition_design|
     lighting_color|story_richness_interest|core_prompt_centrality|sound_design|
     subject_performance_specificity|visual_style|none",
  "primary_issue": "storyboard_timing|camera_design|editing_transition_design|
     lighting_color|story_richness_interest|core_prompt_centrality|sound_design|
     subject_performance_specificity|visual_style|none",
  "summary": "",
  "dimensions": {
    "storyboard_timing": {
      "status": "complete|deficient",
      "severity": "none|minor|major",
      "missing_elements": [],
      "evidence": "",
      "note": ""
    },
    "camera_design": {
      "status": "complete|deficient",
      "severity": "none|minor|major",
```

```
      "missing_elements": [],
      "evidence": "",
      "note": ""
    },
    "editing_transition_design": {
      "status": "complete|deficient",
      "severity": "none|minor|major",
      "missing_elements": [],
      "evidence": "",
      "note": ""
    },
    "lighting_color": {
      "status": "complete|deficient",
      "severity": "none|minor|major",
      "missing_elements": [],
      "evidence": "",
      "note": ""
    },
    "story_richness_interest": {
      "status": "complete|deficient",
      "severity": "none|minor|major",
      "missing_elements": [],
      "evidence": "",
      "note": ""
    },
    "core_prompt_centrality": {
      "status": "complete|deficient",
      "severity": "none|minor|major",
      "missing_elements": [],
      "evidence": "",
      "note": ""
    },
    "sound_design": {
      "status": "complete|deficient",
      "severity": "none|minor|major",
      "missing_elements": [],
      "evidence": "",
      "note": ""
    },
    "subject_performance_specificity": {
      "status": "complete|deficient",
      "severity": "none|minor|major",
      "missing_elements": [],
      "evidence": "",
      "note": ""
    },
    "visual_style": {
      "status": "complete|deficient",
      "severity": "none|minor|major",
      "missing_elements": [],
      "evidence": "",
      "note": ""
    }
  }
}
```

B.3 Representative Case

We provide one representative case together with the outputs from all evaluated models for direct comparison.

Input Prompt

```
A sweet 9-year-old boy stands alone in a quiet neighborhood garden, looking slightly
     downcast. Moments later, his big brother walks up from behind with a kind smile and
     softly wraps him in a comforting embrace. The child's expression instantly brightens
     with joy. Warm afternoon light, vibrant animated style, fluid motion, touching and
     uplifting atmosphere, 3D family-film-style characters.

Duration: 15 seconds.
```

Model outputs.

WanPE-397B

```
The video consists of 4 shots. This 3D Animation, a work in the style of a Disney-Esque 3D
     Animation, is a short scene depicting a touching moment between two brothers. The
     overall art style is Cartoon/Animated, featuring characters with Exaggerated Semi-
     Realistic Proportions, rendered with Stylized PBR materials and illuminated by
     Stylized Realistic Lighting to create a warm, vibrant, and family-friendly atmosphere.
     The story shows a 9-year-old boy who is comforted by his older brother in a
     neighborhood garden. The scene is set in a quiet, sunlit neighborhood garden. A low,
     white picket fence lines the back of the garden, with lush green hedges and blooming
     flowers visible beyond it. A small, leafy tree stands on the left side of the scene.
     The ground is covered in soft, well-kept grass. The scene is bathed in the warm,
     golden light of a late afternoon. The video scene is set at late afternoon, the
     lighting overall led by soft Sunny Light, the image's light-dark contrast is medium-
     low, presenting a bright mid-tone characteristic. Color saturation is high, overlaid
     with a medium-intensity Pale Warm Yellow filter, the whole leaning toward a warm tone
     tendency, together creating a warm and uplifting emotional atmosphere. The narrative
     is presented from a third-person objective perspective using standard cinematic
     animation techniques with a measured editing pace, while the audio landscape combines
     a gentle, evolving piano score and subtle ambient sounds to create a touching and
     uplifting atmosphere.
Shot 1 [00.00-03.50s] The scene opens with a static, eye-level medium shot with a deep
     depth of field, employing a center composition. In the middle of the frame, a 9-year-
     old boy with a sweet face and large eyes stands alone on the soft, green grass. He
     wears a simple blue t-shirt and khaki shorts. The key light source is the soft, warm-
     yellow Sunny Light, forming a front side light effect on the boy and presenting a
     medium lighting ratio. The background features the white picket fence, lush green
     hedges, and blooming flowers, all softly lit by the afternoon sun. The boy performs an
     action of standing downcast. With his shoulders slightly slumped, he looks down at
     the ground, his expression downcast and lonely. His body remains still as a gentle,
     melancholic piano melody begins, accompanied by the faint, ambient sound of birds
     chirping and a soft breeze rustling leaves.
Shot 2 [03.50-06.50s] After a hard cut, the camera shifts to a medium long shot from a low
     angle, with a deep depth of field. The camera then performs a slow, stable dolly-in
     movement, moving forward to focus on the approaching character. The 9-year-old boy
     remains in the midground, his back partially to the camera. From the background, his
     big brother, a teenager with a kind, reassuring smile, wearing a gray hoodie and jeans
     , walks into the frame. The warm, golden sunlight creates a soft rim light on his form
     . The brother performs an action of walking towards the boy. He moves with a calm,
     steady gait, his soft footsteps audible on the grass. As he approaches, the piano
     melody continues, its tone becoming warmer and more hopeful.
Shot 3 [06.50-10.50s] A hard cut transitions to a static, over-the-shoulder medium close-
     up, with a shallow depth of field that focuses on the interaction. The 9-year-old boy'
     s back is in the foreground, while the big brother is in the midground. The brother
     performs an action of embracing the boy. With a gentle, fluid motion, he wraps his
     arms around the boy from behind, the movement accompanied by the soft rustle of
     clothing. He rests his chin on the boy's shoulder, his expression full of warmth. The
     piano melody swells gently, becoming more prominent and emotional, as the ambient
     sounds of birds and breeze continue.
Shot 4 [10.50-15.00s] A final hard cut shifts to a close-up of the boy's face, captured at
     eye level with a shallow depth of field. The camera then performs a slow, stable zoom
     -in, tightening the frame on his expression. The boy is in the center of the frame,
     held in his brother's embrace. The boy's expression changes from downcast to joyful.
     Initially, his face is still, but then his eyes widen slightly, and a bright, genuine
     smile spreads across his face, his expression instantly brightening with joy. The
     camera slowly pushes in, emphasizing this emotional shift. The piano melody reaches a
     warm, uplifting crescendo, holding the note before beginning a slow fade, as the
     ambient sounds of the garden continue.
```

WanPE-35B

```
The video consists of 3 shots. This 3D family-film-style scene, in a warm, vibrant, and
     animated style, depicts a touching moment where a big brother comforts his downcast
     younger brother in a garden. The overall visual style is defined by a 3D Animation
     genre with a 3D family-film-style, featuring characters with a 3D family-film-style,
     rendered with a vibrant, polished, and cinematic quality. The scene is set in a quiet,
     well-kept neighborhood garden during the afternoon. The narrative is presented from a
     third-person objective perspective, using a standard cinematic animation style with a
     slow, deliberate editing pace that emphasizes the emotional connection. The audio
     landscape is composed of a gentle, warm piano score, ambient garden sounds, and a
     single line of soft, comforting dialogue, all contributing to a warm, uplifting, and
     touching atmosphere.
Shot 1 [00.00-04.00s] The scene opens with a medium shot at an eye-level angle, employing
     a center composition. The scene is set in a quiet, well-kept neighborhood garden
     during the afternoon, with a low brick wall and a wooden fence in the background, and
     a lush green lawn in the foreground. The video scene is set at day, the lighting
```

overall led by soft Sunny Light, the image's light-dark contrast is medium-low, presenting a moderate mid-tone characteristic. Color saturation is high, overlaid with a medium intensity Pale Warm Yellow filter, the whole leaning toward a warm tone tendency, together creating a warm and uplifting emotional atmosphere. A 9-year-old boy, with a clean, innocent face, stands alone in the center of the frame, his body facing forward. He wears a simple t-shirt and shorts. The boy is looking slightly downcast. His shoulders are slightly slumped, and his expression is somber as he looks down at the ground, his posture conveying a sense of sadness. The camera performs a very slow, small-amplitude dolly-in, gradually tightening the frame on the boy, which enhances the feeling of his isolation. The key light source is the soft, warm-yellow afternoon sunlight, forming a front side light effect on the boy, presenting a medium lighting ratio. The background is softly lit by the same ambient daylight. The audio landscape is established with a gentle, warm piano melody, accompanied by the soft, ambient sound of a gentle breeze and distant birds chirping, creating a quiet, contemplative mood.
Shot 2 [04.00-09.00s] A hard cut transitions to a static, eye-level medium shot, maintaining the warm, sunny lighting. The 9-year-old boy remains in the center of the frame, still looking downcast. A 15-year-old big brother, with a kind, caring face, walks into the frame from the right, moving behind the boy. He wears a casual t-shirt and jeans. The big brother approaches with a gentle, kind smile, his movements fluid and deliberate. As he reaches the boy, he gently wraps his arms around the boy's shoulders, pulling him into a comforting embrace. The action is smooth and continuous, with the brother's arms moving in a gentle arc. The key light source is the soft, warm-yellow afternoon sunlight, forming a front side light effect on both boys, presenting a medium lighting ratio. The background is softly lit by the same ambient daylight. The piano melody continues, its warmth increasing slightly, and the sound of the brother's soft footsteps on the grass is heard, followed by the gentle rustle of clothing as he embraces the boy.
Shot 3 [09.00-15.00s] A hard cut shifts to a close-up of the 9-year-old boy's face, captured at an eye-level angle. The boy is being embraced by his brother, who is just out of frame. The boy's expression instantly brightens with joy. His eyes, previously downcast, lift to look up towards his brother, and a wide, genuine smile spreads across his face, his eyes crinkling with happiness. The camera performs a slow, small-amplitude dolly-in, emphasizing the boy's joyful reaction. The key light source is the soft, warm-yellow afternoon sunlight, forming a front side light effect on the boy, presenting a medium lighting ratio. The background is softly lit by the same ambient daylight. The piano melody swells to a warm, uplifting crescendo, and the ambient sound of a gentle breeze is present. The big brother speaks, his voice soft and comforting, "It's okay." The sound of a soft, contented sigh from the boy is heard, completing the emotional arc of the scene.

H3-Context-IR

integrated_multimodal_description: [Shot 1] 3D-animated, a medium shot with a slow push in on a 9-year-old boy standing alone in a stylized, vibrant neighborhood garden. The young boy features oversized, expressive hazel eyes, slightly messy chestnut-brown hair, and is dressed in a yellow-and-white striped cotton t-shirt tucked into blue denim overalls with bright red sneakers. He stands quietly on a winding cobblestone path, surrounded by oversized blooming blue hydrangeas and a pristine white picket fence in the background. Warm, golden afternoon sunlight filters through a dense canopy of green leaves above, casting distinct, dappled light rays across his downcast face. His shoulders slump forward as his gaze remains fixed on the ground, his right red sneaker lightly tapping against a small grey pebble. The camera pushes in steadily, emphasizing the boy's drooping posture and melancholic expression. [Shot 2] At 00:07.500, the camera cuts to a medium wide arc shot, circling the pair slowly as the boy's older brother steps into the frame from the background. The brother, a tall teenager with neatly parted brown hair, wears a thick forest-green hoodie and khaki trousers. He approaches the young boy from behind with a gentle, wide smile on his face. Upon reaching the child, the brother drops to one knee and softly wraps both arms around the boy's shoulders in a tight, comforting embrace. As the camera completes its arc to fully reveal the young boy's face, his expression instantly transforms. His large eyes widen, his eyebrows raise, and a massive, toothy smile breaks across his face. The golden afternoon light catches the edges of the brother's green hoodie and the boy's brown hair, creating a bright, glowing rim light around them as they hold the embrace.
overall_soundscape: A continuous ambient track of a gentle breeze and light bird chirps fills the background. In the foreground, a distinct, repetitive scrape and tap is clearly heard as the boy's sneaker hits the cobblestone and the pebble. This is followed by the audible, rhythmic thud of the older brother's footsteps approaching on the stone path, which transitions into a pronounced, loud rustle of heavy cotton fabric as the brother wraps his arms around the boy.
non_diegetic_music: Solo acoustic guitar accompanied by a gentle glockenspiel, slow tempo. The instrumentation begins with sparse, sustained notes, gradually building in volume and shifting to a bright, major-key chord progression in the second half, with no percussion.

LTX-2.5-PE

A medium shot frames a sweet 9-years-old boy with light brown hair and a pale blue t-shirt standing alone in a quiet, lush neighborhood garden filled with vibrant green foliage and colorful flowers, his gaze directed slightly downward, captured from a slightly low-angle perspective under warm afternoon light that casts soft shadows, the scene rendered in a vibrant animated style with fluid motion and rich, saturated film-grade colors, accompanied by a soft, gentle piano melody, a moment later, a tall, smiling big brother with dark hair and a casual light-colored shirt walks up from behind the boy, the camera smoothly tracks backward as the brother approaches, and he softly wraps his arms around the younger boy in a comforting embrace, the camera remains static as the embrace occurs, the boy's downcast expression instantly shifts to a bright, joyful smile, the warm cinematic lighting enhances the texture of the leaves and the soft glow on their faces, creating a touching and uplifting atmosphere with 3D family film-style characters and a subtle swell in the background music.

LingBot-Video-PE

{"comprehensive_description": {"scene_content_description": "In a lush, sun-drenched garden, a young boy with short brown hair and a white t-shirt stands alone, looking down with a somber expression. The background is filled with vibrant green foliage and a large, leafy tree. As the scene progresses, an older boy with curly brown hair, wearing a light blue t-shirt, enters from the right and approaches the younger boy. He gently places his hands on the younger boy's shoulders and then wraps his arms around him in a warm, comforting hug. The younger boy's expression shifts from sadness to a bright, joyful smile as he leans into the embrace. The lighting is warm and golden, suggesting late afternoon, creating a tender and emotional atmosphere.",
"camera_movement_description": "The camera remains essentially stationary throughout the video, maintaining a medium shot that focuses on the two boys. There is a very slight, natural handheld tremor, but no intentional panning, tilting, or zooming occurs."},
"camera_info": {"color": "Warm", "frame_size": "Medium", "shot_type_angle": "Low angle", "lens_size": "Medium", "composition": "Center", "lighting": "Hard light", "lighting_type": "Daylight"},
"world_knowledge": [],
"prominent_elements": [{"name": "young boy", "description": "A young boy with short brown hair and a white t-shirt, appearing sad at the start of the video.", "actions": [{"timestamp": "[0.0s - 4.5s]", "action": "stands still, looking down with a sad expression"}, {"timestamp": "[4.5s - 10.5s]", "action": "remains still as the older boy approaches and hugs him"}, {"timestamp": "[10.5s - 13.0s]", "action": "smiles broadly and leans into the hug"}, {"timestamp": "[13.0s - 15.0s]", "action": "stands still, smiling while being hugged"}], "location": "center of the frame", "relative_size": "medium", "shape_and_color": "humanoid shape, wearing a white t-shirt", "texture": "soft skin, cotton t-shirt", "appearance_details": "short brown hair, white t-shirt with a small graphic on the chest", "relationship": "being comforted by the older boy", "orientation": "facing forward, slightly tilted down", "pose": "standing upright, then leaning into a hug", "expression": "changes from sad and downcast to happy and smiling", "clothing": "white t-shirt", "gender": "male", "skin_tone_and_texture": "fair, smooth skin"},
{"name": "older boy", "description": "An older boy with curly brown hair and a light blue t-shirt, acting as a comforting figure.", "actions": [{"timestamp": "[4.5s - 7.0s]", "action": "enters from the right and walks toward the younger boy"}, {"timestamp": "[7.0s - 10.5s]", "action": "places hands on the younger boy's shoulders and hugs him"}, {"timestamp": "[10.5s - 15.0s]", "action": "holds the younger boy in a hug, smiling"}], "location": "enters from the right, moves to the center", "relative_size": "medium", "shape_and_color": "humanoid shape, wearing a light blue t-shirt", "texture": "curly hair, cotton t-shirt", "appearance_details": "curly brown hair, light blue t-shirt", "relationship": "comforting the younger boy", "orientation": "facing left toward the younger boy", "pose": "walking, then standing and hugging", "expression": "kind and smiling", "clothing": "light blue t-shirt", "gender": "male", "skin_tone_and_texture": "fair, smooth skin"}]}

SCMaPR

A sweet 9-year-old boy stands alone in a quiet neighborhood garden, looking slightly downcast with a slightly downcast expression under warm afternoon light. His older brother walks up gently from behind and softly wraps his arms around the younger boy in a comforting embrace. The younger boy's face immediately brightens with joy in direct response to this tender physical contact. The scene is rendered in a vibrant animated style with fluid motion and 3D family-film-style characters to create a touching and uplifting atmosphere.

VPO

A sweet 9-year-old boy with tousled hair and a plain T-shirt stands alone in a quiet neighborhood garden, surrounded by blooming flowers and a wooden bench under the warm sunlight. His posture is slightly slouched, and his gaze is cast downward, suggesting a moment of quiet melancholy. The scene captures the soft rustle of leaves and the gentle sway of nearby trees under the golden afternoon light, adding to the serene yet slightly somber mood. Moments later, his big brother, slightly taller and wearing a casual hoodie, walks up from behind with a warm and kind smile. The boy's expression softens as his older brother approaches, and with a heartfelt gesture, he gently wraps his arms around him in a comforting embrace. The younger boy's face lights up with pure joy, radiating warmth and the unspoken bond of sibling love. The animated style features soft, fluid motion and vibrant colors, enhancing the heartfelt and uplifting atmosphere of the moment. The background includes subtle details like a colorful butterfly fluttering near the flowers and distant chirping birds, reinforcing the wholesome, family-film-style aesthetic.

Prompt-A-Video-Cog

In a serene 3D neighborhood garden, a cherubic 9-year-old boy with a softly tousled hairdonned a lightweight jacket, standing alone under the embracing branches of a oak tree. His eyes, cast downward, whisper stories of a gentle sorrow. As his stalwart big brother, with a warm smile, wraps him in a comforting embrace, the younger boy's expression transforms, brightening with joy. The brothers enclave is bathed in the warm, honeyed light of a late afternoon sun. The animation is vibrant, with fluid motions, captivating in a style reminiscent of timeless family films. The scene is a powerful testament to the embracing love of a big brother, as the animation exudes a touching, uplifting atmosphere that lasts for 15 captivating seconds.

Prompt-A-Video-OS

In a vibrant neighborhood garden bathed in warm afternoon light, a 9-year-old boy with a downcast expression sits alone. His big brother, with a kind face, walks into the frame and envelops him in a comforting hug. The boy's demeanor instantly brightens with joy. The video is rendered in a 3D animated style, characteristic of family films, and exudes a touching and uplifting atmosphere. (15 seconds)

RAPO

a young boy wearing a red shirt and sitting on the ground. He is then seen being hugged by his big brother. The scene is set in a warm and sunny afternoon. The animation is colorful and vibrant. The boy's expression changes from sadness to joy as he is being hugged by his big brother.

C Additional Experimental Details

C.1 Implementation Details

We evaluate the ten prompt-enhancement systems listed in Table 2, together with the original prompt as a no-enhancement baseline. All systems are evaluated on T2V-PE, while systems supporting visual inputs are additionally evaluated on I2V-PE and R2V-PE. Proprietary systems are accessed through their official interfaces, and released models are run using their recommended inference settings.

For each case, modality-aware evidence is extracted once and cached for all methods. GPT-5.4 (OpenAI, 2026a) extracts facts for image-only inputs, and Qwen3.5-Omni-Plus processes all assets for cases containing video references. GPT-5.4 then evaluates every enhanced prompt using the same criterion-specific rubrics. We report criterion pass rates and aggregate Semantic Fidelity, Internal Consistency, and Directorial Creation scores. The additional experiments analyze temporal complexity, reference complexity, and output length. Human studies further test prompt-level alignment and downstream utility using videos generated by Wan3.0 and MiniMax-H3. For downstream video generation, we access Wan3.0 through its online API with prompt enhancement disabled, while MiniMax-H3 is run locally on eight NVIDIA A100 GPUs.

C.2 Length-Stratified Analysis

Prompt enhancement often transforms concise inputs into substantially longer descriptions. To test whether output length alone yields better descriptive adequacy, we analyze H3-Context-IR separately on each task. Valid outputs are sorted by token length and divided into four equal-sized quartiles, from the shortest ($L1$) to the longest ($L4$).

Table 8: **Length-stratified results for H3-Context-IR.** Quartiles are defined independently within each task. DC denotes Directorial Creation.

Task	Metric	$L1$	$L2$	$L3$	$L4$
T2V-PE	Avg. tokens	394.7	490.9	584.9	861.1
	DC	79.32	78.60	77.35	74.13
I2V-PE	Avg. tokens	819.1	945.7	1066.6	1382.0
	DC	75.00	75.20	75.99	67.20
R2V-PE	Avg. tokens	809.9	1015.7	1164.4	1479.3
	DC	79.37	73.75	72.07	69.28

Across all three tasks, the longest quartile underperforms the shortest by 5.19 to 10.09 DC points. Thus, longer outputs do not consistently provide better directorial guidance. This result is consistent with the evaluation design: DC rewards concrete, relevant, and executable specifications for temporal planning, cinematography, narrative progression, subject performance, and sound, rather than verbosity or repeated template language.

D Human Expert Annotation

D.1 Pairwise Annotation Protocol

The human study evaluates alignment at two levels: (i) the quality of enhanced prompts and (ii) the quality of downstream videos generated from those prompts. We recruit 60 experts and provide task-specific training before annotation. Each expert evaluates approximately 50 T2V-PE pairs, 20 I2V-PE pairs, and 10 R2V-PE pairs. The comparison pool contains 11 systems for T2V-PE, 4 for I2V-PE, and 3 for R2V-PE, including the original prompt as the no-enhancement baseline in every task.

Pairwise judgment procedure.

For enhanced-prompt evaluation, annotators inspect the original instruction, target duration, reference assets when applicable, and an anonymized pair of enhanced prompts. Annotators make a separate choice for Semantic Fidelity, Internal Consistency, and Directorial Creation. For each dimension, they select the better candidate, both good, or both bad. For downstream-video evaluation, annotators inspect the source request and references together with an anonymized pair of generated videos and make the same dimension-specific judgment. Candidate identities are hidden, and the left and right positions are randomized independently for every comparison.

We aggregate the pairwise choices with a Bradley-Terry model to obtain a system ranking for each task, evaluation level, and dimension. We then compute the Spearman correlation between this expert-derived ranking and the corresponding automatic dimension scores. The resulting coefficients are reported in Table 5.

Quality control and agreement.

All experts complete training before formal annotation. We test each annotator's selections for a significant left or right position preference. Annotators exhibiting significant position bias are excluded, together with all of their judgments, before fitting the Bradley–Terry models and computing the correlations.

D.2 Annotation Interface

The interface presents the original user prompt and target duration at the top, followed by reference assets for I2V-PE and R2V-PE. Prompt-level evaluation displays two anonymized enhanced prompts, while video-level evaluation replaces them with two generated videos. The three evaluation dimensions are shown as independent rating rows.

Interface behavior. Candidate identities are replaced by A and B, and their left and right positions are randomized independently for each case. Video-level evaluation supports synchronized playback, individual playback, and replay. The interface requires one response for each dimension before submission and allows annotators to revise their selections before completing the case. Figures 5–7 show the annotation interfaces for the three PE settings.

E Ethics, Privacy, and Licensing

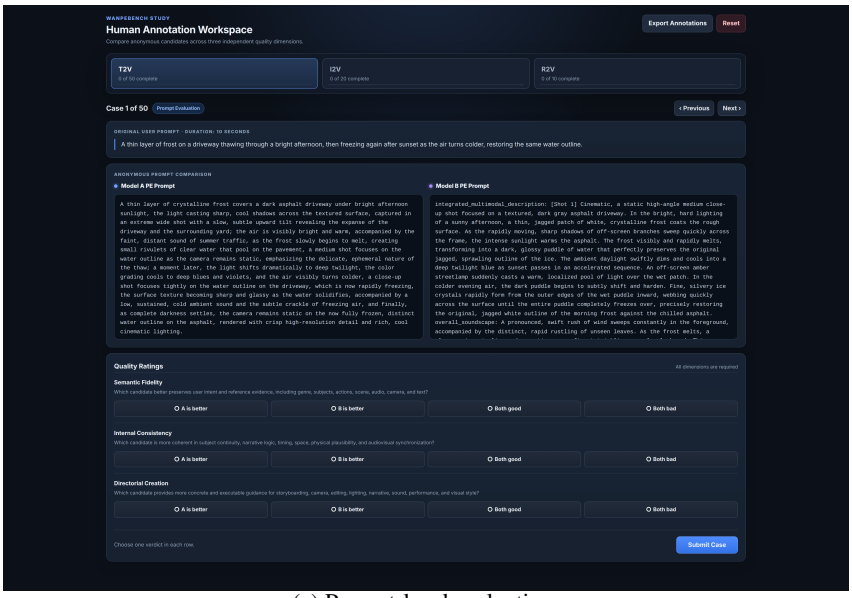
WanPEBench prompts are generated from an expert-designed taxonomy and undergo manual verification. During curation, we remove personal information, identifiable private individuals, sensitive political content, and material that may raise privacy or cultural concerns. Reference assets originate from public benchmarks or model-assisted generation. We retain source and license metadata and release only assets permitted for redistribution, providing source identifiers for restricted material. Videos generated for downstream validation are used only for evaluation and are not included in the benchmark release. We release the verified prompts, legally shareable reference assets, and evaluation framework to support reproducible research.

F Limitations

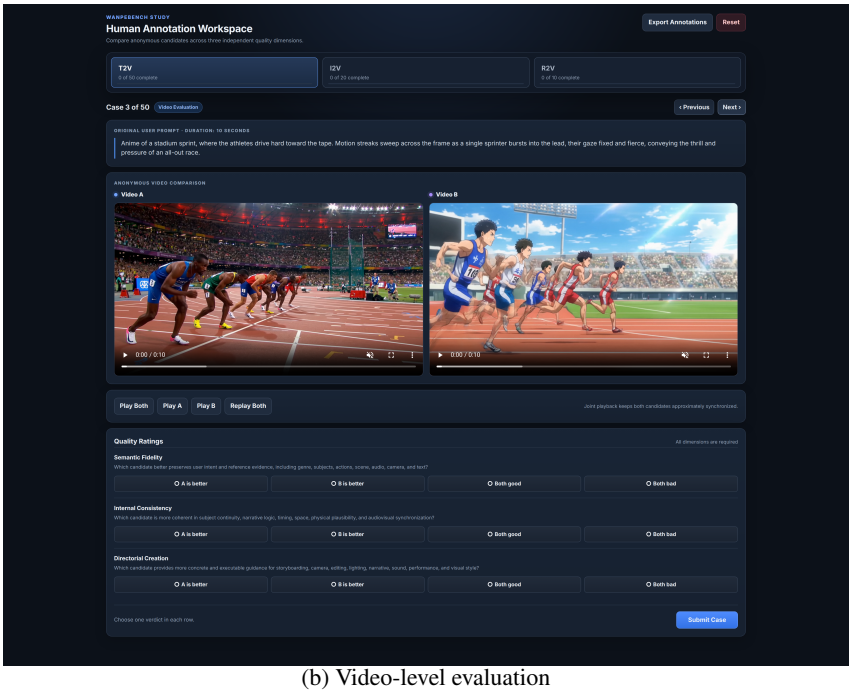
One limitation is the scale of the benchmark. Although 1,100 expert-verified cases provide broad coverage across three PE settings, 35 task types, and diverse conditioning patterns, they cannot represent the full range of rapidly evolving video-generation requests. Future work will expand WanPEBench with additional domains, languages, reference configurations, and emerging control formats, enabling more comprehensive evaluation and supporting the development of increasingly general prompt enhancers.

A second limitation is the comparatively limited model coverage for I2V-PE and R2V-PE. Unlike T2V-PE, few existing methods support multimodal prompt enhancement, and only a small number of proprietary systems, such as H3-Context-IR, provide accessible PE interfaces. Consequently, our I2V-PE and R2V-PE evaluations include fewer baselines. We will continue updating WanPEBench as more multimodal PE methods become available, enabling broader comparisons across open and proprietary systems.

T2V-PE Annotation Interface



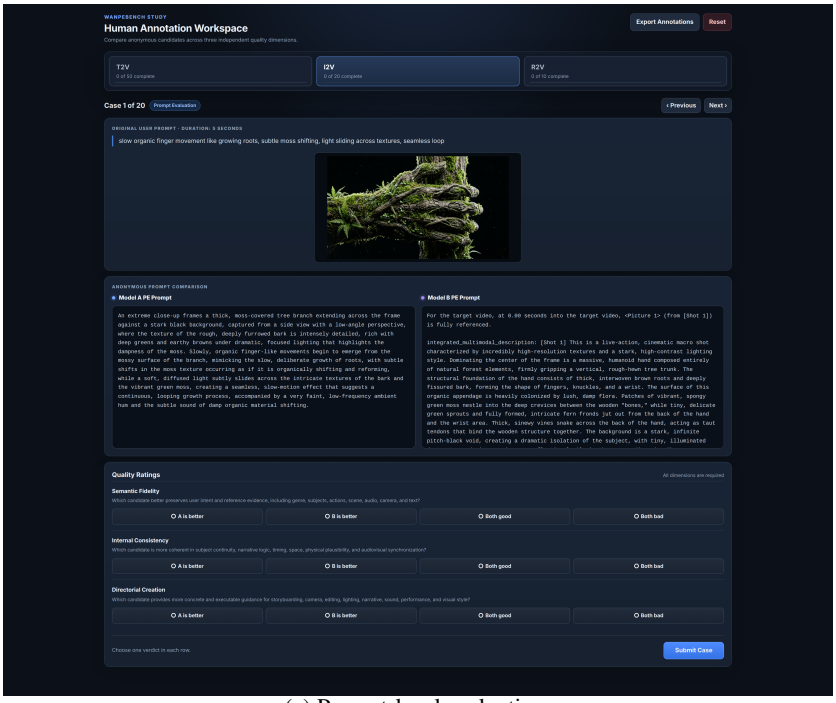
(a) Prompt-level evaluation



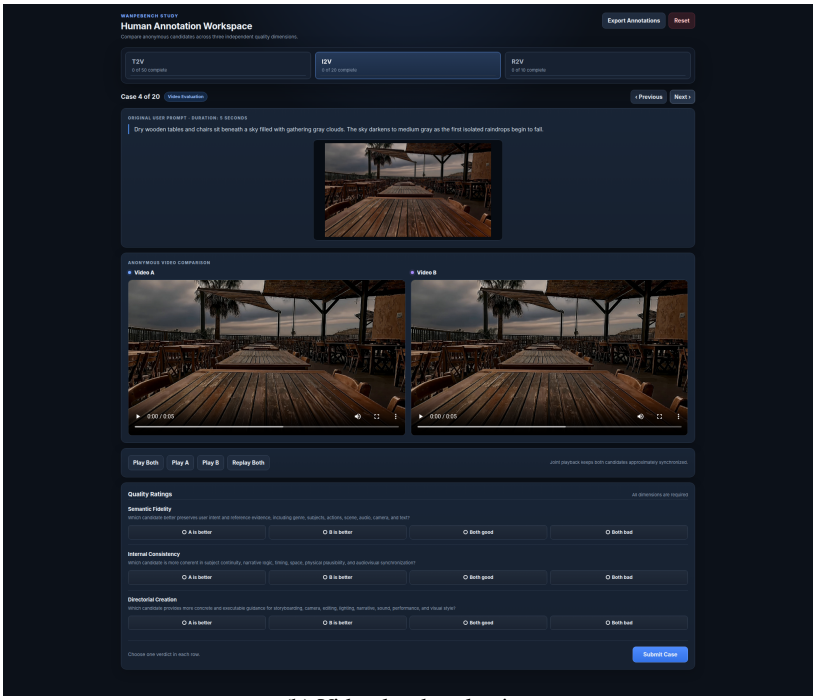
(b) Video-level evaluation

Figure 5: **T2V-PE annotation interface.** Anonymous prompts and generated videos are compared against the text instruction and target duration.

I2V-PE Annotation Interface



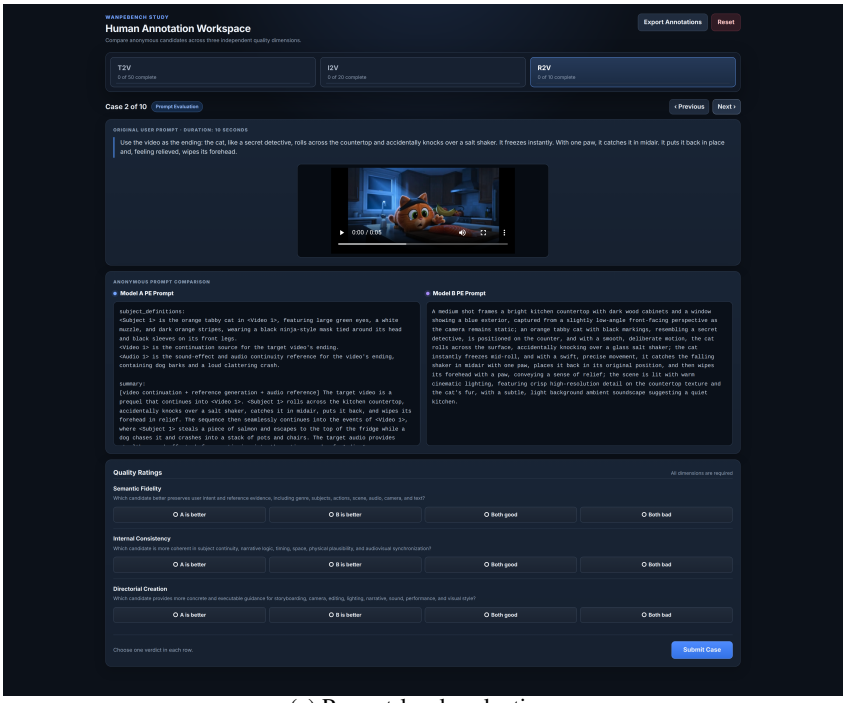
(a) Prompt-level evaluation



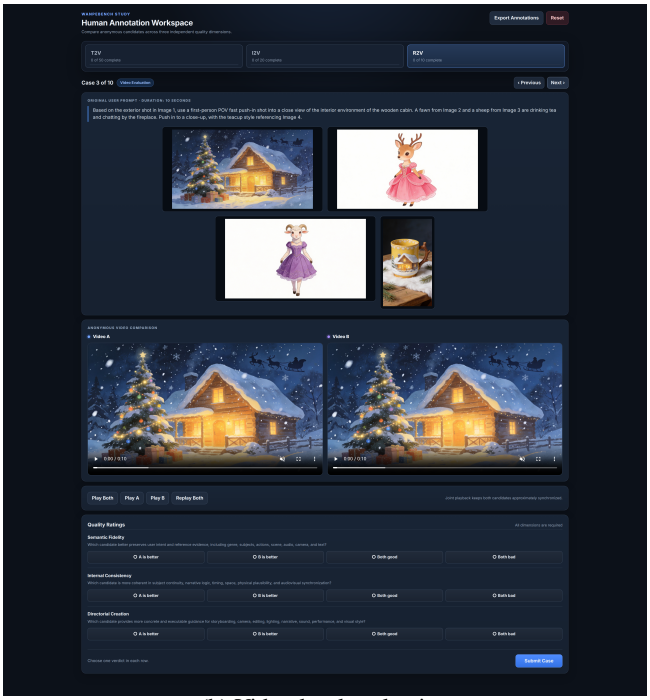
(b) Video-level evaluation

Figure 6: **I2V-PE annotation interface.** The initial-frame image is displayed with the instruction for prompt-level and video-level comparison.

R2V-PE Annotation Interface



(a) Prompt-level evaluation



(b) Video-level evaluation

Figure 7: **R2V-PE annotation interface.** All reference images and videos are displayed to support cross-asset grounding judgments.