RVEBENCH: A COMPREHENSIVE BENCHMARK FOR REFERENCE-GUIDED VIDEO EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reference-guided video editing is evolving toward complex workflows that combine natural-language instructions with multiple visual references. However, existing benchmarks do not systematically construct challenge-specific cases targeting advanced capabilities and therefore fail to expose capability-specific failures. To address this gap, we introduce RVEBench, a comprehensive and diagnostic benchmark comprising 400 human-calibrated cases and 5,672 fine-grained checklist questions, with six challenging scenarios spanning Spatial Perception, Temporal Modeling, Semantic Reasoning, and Symbolic Fidelity. We further introduce RVE-Eval, a five-dimensional framework that combines sample-specific checklists and rubrics to distinguish edit execution, reference grounding, source preservation, perceptual quality, and physical or logical plausibility. Experiments on ten configurations spanning seven representative models reveal several fundamental limitations. Current systems often produce perceptually polished videos that remain physically or logically implausible. Challenging scenarios expose distinct capability-specific failure patterns. As editing requirements accumulate, instruction execution remains stable, but unintended changes and quality degradation become increasingly severe.

1 INTRODUCTION

Video editing is shifting from text-only manipulation toward multimodal, reference-guided control. Natural-language instructions specify desired changes but often underspecify fine-grained appearance such as identity, clothing, texture, material, or style (Cheng et al., 2024; Jiang et al., 2025; Bai et al., 2026; He et al., 2026; Rao et al., 2026), whereas reference images provide direct visual anchors (Lin et al., 2026; Bernini Team et al., 2026; Wang et al., 2026a). As recent systems begin to accept multiple references and coordinate several editing requirements (Wei et al., 2026a; Bernini Team et al., 2026; Liu et al., 2026b; MiniMax, 2026a), the key question is whether they can reliably ground and execute complex reference-guided edits throughout a video.

Models may complete direct replacements yet fail when targets are ambiguous or inconspicuous, entities move rapidly or recur across shots, edits require world knowledge, or reference text must be reproduced exactly. Figure 1 organizes these challenges into four capabilities beyond routine visual transfer: *Spatial Perception* for target identification, *Temporal Modeling* for correspondence over time, *Semantic Reasoning* for unstated constraints, and *Symbolic Fidelity* for exact symbolic reproduction.

However, existing benchmarks do not expose these capability-specific failures. Instruction-guided benchmarks assess instruction adherence, source preservation, and visual quality without evaluating visual-reference grounding (Sun et al., 2025; Chen et al., 2026b; Wu et al., 2026). Reference-guided benchmarks cover limited edit dimensions or only single-reference settings (Lin et al., 2026; Bernini Team et al., 2026; Pan et al., 2026; Miao et al., 2026); UniVBench and TIDE-Bench support multiple references but lack systematically constructed cases targeting advanced capabilities (Wei et al., 2026b; Liu et al., 2026a). Capability coverage alone cannot distinguish failures in edit execution, reference grounding, source preservation, perceptual quality, or scene plausibility. Diagnostic evaluation requires challenge-specific cases and dimensions that separate these failure sources.

To address this gap, we introduce **RVEBench**, a capability-centered benchmark comprising 400 human-calibrated cases, each with one to four reference-grounded editing requirements, and 5,672

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 RVEBench.** Representative tasks from Basic Editing and six challenging scenarios organized around four capabilities. Parentheses indicate numbers of test cases.

checklist questions. It includes 160 *Basic Editing* cases and 240 *Advanced* cases across six scenarios probing four capabilities, separating routine competence from capability-specific degradation.

To diagnose failure sources, we develop **RVE-Eval**, which separates editing success into five dimensions: *Instruction Following*, *Reference Consistency*, *Video Consistency*, *Video Quality*, and *Video Realism*. It uses sample-specific checklists for the three condition-dependent dimensions and rubrics for the two intrinsic dimensions, Video Quality and Video Realism.

Evaluating ten configurations across seven models reveals that physical and logical plausibility lags behind perceptual quality, so polished videos may remain implausible. Advanced scenarios expose capability-specific degradation, while increasing edit counts primarily harms source preservation and perceptual quality rather than edit execution. Prompt enhancement improves editing effectiveness more than generation quality, and neither cascaded single-reference editing nor combining multiple references into a single grid image reliably substitutes for native multi-reference editing.

Our main contributions are threefold:

- We introduce **RVEBench**, comprising 400 human-calibrated cases spanning Basic Editing and six Advanced scenarios that probe four capabilities under one to four reference-grounded requirements.

- We develop **RVE-Eval**, which combines checklists and rubrics across five dimensions to distinguish edit execution, reference grounding, source preservation, perceptual quality, and physical or logical plausibility.
- We provide capability- and dimension-level analyses that expose systematic degradation patterns and failure modes in reference-guided video editing.

2 Related Work

Advancements in Video Editing Models. Video editing has progressed from text-driven manipulation to multimodal, reference-guided control. Early approaches extended pretrained image diffusion models to video by introducing temporal attention, optical flow, or inversion-based mechanisms (Wu et al., 2023; Liang et al., 2024; Qi et al., 2023), but they often exhibited limited temporal consistency and editing flexibility. Subsequent work shifted toward instruction-driven models (Cheng et al., 2024; He et al., 2026; Rao et al., 2026), and unified frameworks, such as VACE (Jiang et al., 2025), that accommodate diverse editing conditions within a shared architecture. More recent systems integrate reference images with MLLM-based semantic planning to specify visual details that are difficult to express precisely through text (Wei et al., 2026a; Chen et al., 2026a; Lin et al., 2026; Bernini Team et al., 2026). Emerging methods further explore agentic planning for underspecified requests (Yu et al., 2026) and explicit binding for coordinating multiple references (Liu et al., 2026b).

Video Editing Benchmarks. Video editing benchmarks have advanced rapidly, yet complex reference-guided scenarios remain underexplored. VE-Bench (Sun et al., 2025) and IVEBench (Chen et al., 2026b) primarily assess instruction-guided editing based on instruction adherence, source preservation, and visual quality. More recent benchmarks incorporate visual references, but their coverage remains restricted: RefVIE-Bench (Lin et al., 2026) concentrates on subject reference and background replacement, whereas Bernini-Bench (Bernini Team et al., 2026) covers a broader range of editing dimensions but remains restricted to a single reference image per sample. UniVBench (Wei et al., 2026b) and TIDE-Bench (Liu et al., 2026a) support multiple reference images but lack systematically constructed cases targeting advanced capabilities.

3 RVEBench

3.1 Dataset Design

To provide a diagnostic evaluation, RVEBench organizes test cases into seven scenario categories: *Basic Editing* assesses models' *General Editing* capability, while six challenging scenarios probe four key advanced capabilities—*Spatial Perception*, *Temporal Modeling*, *Semantic Reasoning*, and *Symbolic Fidelity*.

◇ **General Editing.** Forms the foundation for reference-guided video editing across diverse edit dimensions. *Basic Editing* tests whether models can correctly interpret and execute editing requirements grounded in corresponding reference images.

◇ **Spatial Perception.** Evaluates whether models can accurately identify and edit intended targets in complex visual scenes. *Target Disambiguation* tests whether models can resolve the intended target among similar candidates using appearance, spatial-relation, behavioral, or semantic-role cues while avoiding unintended changes to other entities, whereas *Inconspicuous Entities* tests the detection and editing of small, peripheral, low-saliency, or partially occluded targets.

◇ **Temporal Modeling.** Temporal challenges arise when edits must either remain aligned with dynamic content or respect the shot structure of the source video. *Fast Movement* challenges target binding under rapid motion, articulation, deformation, or interactions among multiple entities, where models are prone to action misalignment, identity drift, and temporal flicker. *Multi-Shot Editing* evaluates entity association across shots, requiring edits to propagate consistently to all relevant shots while remaining absent from unrelated shots.

◇ **Semantic Reasoning.** Semantic reasoning is required when visual correspondence alone is insufficient. *World Knowledge* tests whether models can produce contextually appropriate outcomes that respect physical laws, reflect identity and cultural knowledge, or adapt to the surrounding environment.

◇ **Symbolic Fidelity.** Symbolic fidelity requires discrete visual symbols to be generated or preserved precisely. *Text Fidelity* evaluates the accurate generation and preservation of textual content through text insertion, text replacement, and source-text preservation, focusing on character correctness, font fidelity, and layout consistency.

3.2 Dataset Construction

RVEBench's evaluation framework imposes stringent requirements on sample quality. To satisfy these requirements, we employ a comprehensive and rigorous data annotation and quality control process, as illustrated in Figure 2. This process yields a dataset comprising 400 videos with one to four reference-grounded editing requirements per video and 5,672 fine-grained checklist questions. Additional dataset statistics are provided in Appendix C.3.

Media Diversity via a Hierarchical Taxonomy. To ensure broad coverage of realistic editing scenarios, we construct a diverse media pool of source videos and reference images. We organize videos using a two-level taxonomy with six top-level domains and 27 fine-grained categories (Appendix C.1), collecting them from Pexels, YouTube, and Bilibili. We classify reference images into 16 categories (Appendix C.2) by intended editing purpose (e.g., humans, animals, clothing, backgrounds, makeup, colors, and styles), sourcing them from Pexels, Google web image search, and GPT-Image-2 synthesis (OpenAI, 2026). Together, these taxonomies ensure diversity in source content and reference conditions. Five annotators then annotate the collected media with reference-grounded editing instructions.

MLLM-Based Checklist Generation. To support fine-grained evaluation of editing outcomes, we use Gemini-3.1-Pro-Preview (Gemini Team, 2026), Gemini-3.5-Flash (Google DeepMind, 2026), and Qwen3.8-Max (Qwen Team, 2026b) to generate sample-specific checklists. Given the source video, the editing instruction, all reference images, and the designated evaluation dimension, the models produce detailed and verifiable binary questions that assess individual editing requirements or the sample as a whole. Each question is stated in positive form and assigned an importance weight. The resulting checklists enable systematic evaluation of instruction-following accuracy, consistency with the reference images, and preservation of the source video. The checklist-generation prompts are provided in Appendix H.1.

Criteria-Guided Human Calibration. Three independent quality-control reviewers, who were not involved in the original data collection or annotation stages, conduct the final human calibration of each sample and its associated checklist. Reviewers first verify that each editing instruction clearly identifies both the referenced visual content and the editing target in the source video. All instructions follow the Basic Format, which specifies only the editing target, operation, and reference source without describing the complete expected output. Reviewers then confirm that the editing requirements are mutually consistent and physically feasible. Finally, they ensure that each checklist question is objective, verifiable, and restricted to its designated evaluation dimension. Each sample is independently reviewed by at least two of the three reviewers. This criteria-guided calibration process improves the validity of the dataset and the reliability of the evaluation.

3.3 RVE-Eval

3.3.1 Evaluation Dimensions

As illustrated in Figure 2, we construct a comprehensive, structured, and interpretable evaluation framework comprising the following dimensions:

◇ **Instruction Following.** This dimension evaluates whether each instruction is correctly executed on the intended target, including the requested operation, outcome, and explicit spatial, quantitative, temporal, or action constraints.

◇ **Reference Consistency.** This dimension measures whether the edited content preserves the visual anchors invoked from the corresponding reference image, such as identity, shape, color, texture, clothing, or style.

◇ **Video Consistency.** This dimension evaluates editing fidelity to the source video: content outside the permitted changes should remain stable. It considers non-target entities, unspecified target attributes, actions, scene structure, and camera or temporal organization.

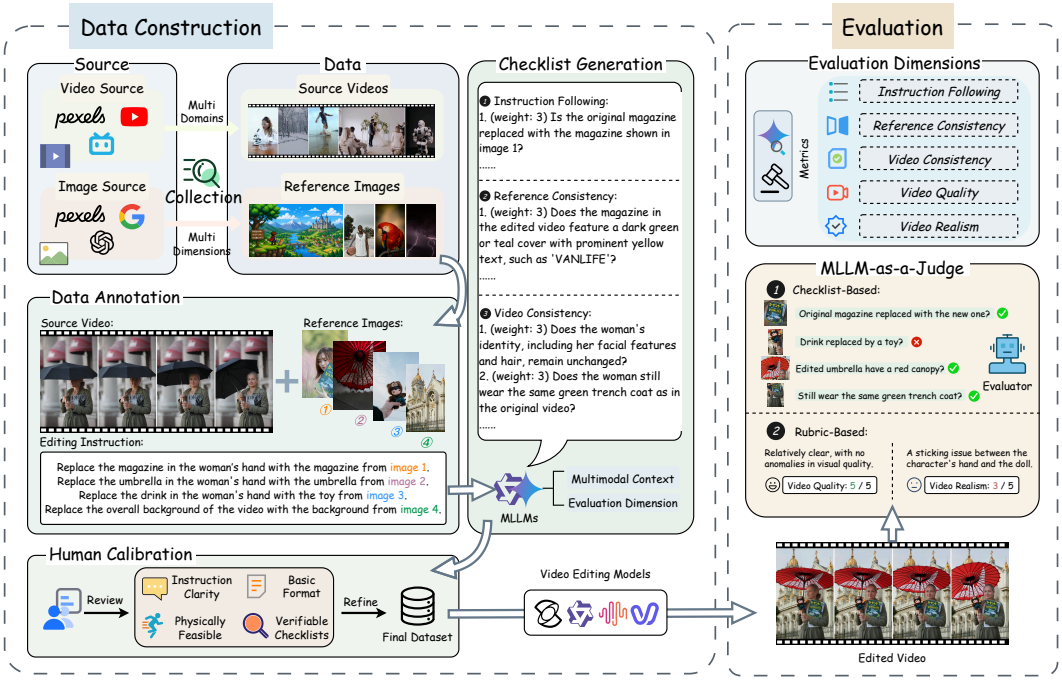


Figure 2: Overview of the RVEBench dataset construction pipeline and RVE-Eval framework.

◇ **Video Quality.** This dimension assesses the technical and perceptual integrity of the edited video, including spatial clarity, local artifacts, temporal stability, visibility, and playback completeness.

◇ **Video Realism.** This dimension assesses whether the edited video remains physically and logically plausible within its established visual style, considering morphology, motion and interactions, spatial relationships, temporal causality, and scene integration.

3.3.2 Evaluation Methodology

RVE-Eval employs two complementary evaluation protocols, using Gemini-3.1-Pro-Preview (Gemini Team, 2026) as the MLLM evaluator. We adopt rubric-based evaluation for **Video Quality** and **Video Realism**, and checklist-based evaluation for **Instruction Following**, **Reference Consistency**, and **Video Consistency**. Together, these protocols assess both the intrinsic properties of the edited video and its compliance with the editing conditions. The corresponding evaluation prompts and scoring rubrics are provided in Appendix H.2.

Rubric-based Evaluation. For Video Quality and Video Realism, the evaluator observes only the video and assigns a case-level integer score from 1 to 5 according to a fixed dimension-specific rubric. We use each source video as a scoring anchor for the corresponding case. Given N cases, the dataset-level score for dimension d is computed as

$$S_d = \frac{100}{N} \sum_{i=1}^{N} \begin{cases} 1, & s_{i,d}^{\mathrm{src}} = 1, \\ \min\left(\dfrac{(s_{i,d}^{\mathrm{edit}} - 1)/4}{(s_{i,d}^{\mathrm{src}} - 1)/4}, 1 \right), & s_{i,d}^{\mathrm{src}} > 1. \end{cases} \tag{1}$$

where $s_{i,d}^{\mathrm{src}}$ and $s_{i,d}^{\mathrm{edit}}$ denote the source- and edited-video scores for case i, respectively. The reasons for using this source-anchored design are discussed in Appendix D.

Checklist-based Evaluation. Individual edits are evaluation units for Instruction Following and Reference Consistency, whereas complete cases are used for Video Consistency. For each unit u and dimension d, the evaluator receives the source and edited videos, instructions, reference images, and the corresponding checklist, then returns a binary judgment for each question. The weighted unit score is

$$C_{u,d} = \frac{\sum_j w_{u,d,j} y_{u,d,j}}{\sum_j w_{u,d,j}} \times 100, \tag{2}$$

Table 1: Comparison between RVEBench and related reference-guided video editing benchmarks. For unified benchmarks, statistics are reported only for their reference-guided video editing subsets. "-" denotes unreported statistics. CA, RB, and CB denote conventional automatic, rubric-based, and checklist-based evaluation, respectively.

Benchmark	#Samples	#Ref. Images	Avg. Images	#Edit Dimensions	Spatial Perception	Temporal Modeling	Semantic Reasoning	Symbolic Fidelity	Evaluation Methods
RefVIE-Bench (Lin et al., 2026)	110	1	1	4	✗	✗	✗	✗	RB
Goku-Bench (Liang et al., 2026)	200	1	1	4	✗	✗	✗	✗	CA RB
IntelligentVBench (Pan et al., 2026)	210	1	1	4	✗	✗	✗	✗	RB
Bernini-Bench (Bernini Team et al., 2026)	80	1	1	6	✗	✗	✗	✗	RB
OmniEdit-Bench (Miao et al., 2026)	65	1	1	7	✗	✗	✗	✓	RB
VicEdit-Bench (Wang et al., 2026b)	100	1/2	1.7	6	✗	✗	✗	✗	CA RB
TIDE-Bench (Liu et al., 2026a)	211	1/2/3	-	6	✗	✗	✗	✗	RB
UniVBench (Wei et al., 2026b)	1240	1/2/3	1.53	6	✗	✓	✗	✗	CB
RVEBench (Ours)	**400**	**1/2/3/4**	**2**	**7**	✓	✓	✓	✓	RB CB

where $y_{u,d,j} = 1$ for `yes` and 0 for `no`, and $w_{u,d,j} \in \{1, 2, 3\}$ is the predefined importance weight. The dataset-level score S_d averages $C_{u,d}$ over all edits $\mathcal{E}$ for the edit-level dimensions and over all cases for Video Consistency.

3.4 Dataset Comparison

In Table 1, we compare RVEBench with existing reference-guided video editing benchmarks. While existing benchmarks do not systematically construct cases targeting advanced capabilities, RVEBench evaluates spatial perception, temporal modeling, semantic reasoning, and symbolic fidelity through six challenging scenarios. Under the same RVE-Eval protocol, Kling-3.0-Omni consistently scores lower on RVEBench's Advanced Scenarios than on UniVBench across all five dimensions (Figure 3). Its Instruction Following score drops from 83.98 to 68.31, while Video Realism decreases from 77.50 to 64.79. These results suggest that RVEBench poses greater challenges for reference-guided video editing, particularly in scenarios requiring advanced capabilities.

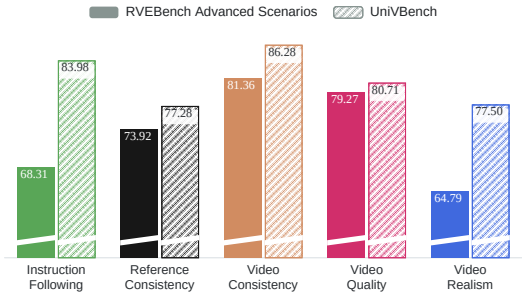


Figure 3: Performance comparison of Kling-3.0-Omni on the RVEBench Advanced Scenarios and UniVBench.

4 Experiment

4.1 Experimental Settings

Evaluated Models. We evaluate 10 configurations spanning 7 representative video editing models that support multi-image reference inputs, comprising four proprietary models and three open-source models. The proprietary models include Kling-3.0-Omni (Kling AI, 2026), Wan2.7-VideoEdit (Alibaba Cloud, 2026), ViduQ2-Pro (Vidu, 2026), and Kling-O1 (Kling Team et al., 2025). The open-source models include MiniMax-H3 (MiniMax, 2026a), Bernini (Bernini Team et al., 2026), and UniVideo (Wei et al., 2026a).

Implementation Details. We access proprietary models through their official inference APIs and, unless restricted by the platform, generate 720p videos matching the source-video duration with their officially designated prompt enhancers enabled. For open-source models, all experiments are conducted on eight NVIDIA H800 GPUs. We follow the officially recommended inference settings, including resolution and frame rate, adjusting only the output duration to match the source video. For prompt enhancement, MiniMax-H3 uses the official H3-Context-IR (MiniMax, 2026b), while Bernini uses Gemini-3-Flash-Preview (Gemini Team, 2025). Detailed inference configurations are provided in Appendix F.

Table 2: Results of 10 configurations spanning 7 models on RVEBench. Basic denotes Basic Editing, whereas Advanced aggregates the remaining six scenarios. IF, RC, VC, VQ, and VR denote Instruction Following, Reference Consistency, Video Consistency, Video Quality, and Video Realism, respectively. Avg. and Overall Avg. denote group-wise and overall averages, respectively. "+" separates the parameter counts of the MLLM and DiT components. indicate the best, second-best, and third-best performance in each dimension, respectively.

Method	Params.	Basic						Advanced						Overall Avg.
		IF	RC	VC	VQ	VR	Avg.	IF	RC	VC	VQ	VR	Avg.	
Proprietary Models														
Kling-3.0-Omni	🔒	87.50	77.70	88.27	81.35	68.54	80.67	68.31	73.92	81.36	79.27	64.79	73.53	76.75
Wan2.7-VideoEdit	🔒	**95.02**	85.21	89.19	70.05	61.82	80.26	81.60	78.67	78.60	64.63	57.26	72.15	75.78
ViduQ2-Pro	🔒	81.64	77.22	77.06	73.47	65.40	74.96	61.91	72.34	60.51	68.89	57.27	64.19	68.89
Kling-O1	🔒	77.58	70.93	82.89	74.48	64.79	74.13	55.87	62.01	71.31	71.35	59.03	63.91	68.49
Open-source Models														
MiniMax-H3	33B	91.91	**88.07**	**89.47**	**84.43**	**71.72**	**85.12**	**89.22**	**86.91**	87.23	**87.68**	**70.19**	**84.24**	**84.66**
Bernini	7B+14B	92.05	84.05	85.02	72.50	66.82	80.09	76.23	73.41	82.66	75.63	63.30	74.24	77.00
Bernini-R	14B	89.14	81.81	87.56	67.14	61.46	77.42	76.00	70.44	**91.20**	76.28	62.36	75.26	76.51
Bernini-R	1.3B	82.94	74.26	77.91	62.92	58.18	71.24	63.45	63.33	82.08	62.99	53.68	65.10	68.04
UniVideo (hidden)	7B+13B	35.19	30.54	78.10	62.08	61.04	53.39	27.03	33.34	60.43	56.60	56.18	46.72	49.47
UniVideo (query)	7B+13B	28.75	24.67	82.32	63.39	61.93	52.21	23.71	28.15	64.42	58.47	53.61	45.67	48.31

4.2 Main Results

Top Open-Source Models Match or Outperform Proprietary Counterparts. Table 2 shows that MiniMax-H3 leads in most evaluation dimensions and achieves the highest overall average score of 84.66 across all evaluated configurations, while Bernini performs comparably to the strongest proprietary model, Kling-3.0-Omni. Together, these results show that leading open-source models can rival proprietary counterparts.

Challenging Scenarios Remain Difficult. Compared with the Basic setting, the Advanced setting yields a 6.45-point average decline and widens the across-configuration score range from 32.91 to 38.57. This added degradation stems mainly from weaker edit understanding and control, not generation quality: Instruction Following and Video Consistency fall by 13.84 and 7.80 points, respectively, versus only 1.00 point for Video Quality. Robustness also varies markedly across configurations—MiniMax-H3 drops just 0.88 points and ranks first in both settings, whereas ViduQ2-Pro shows the largest decline, at 10.77 points.

Video Realism Remains a Major Bottleneck. Across all reported model configurations, Video Realism is lower than Video Quality in both settings; even for MiniMax-H3, the strongest overall model, the gaps are 12.71 points in Basic (84.43 vs. 71.72) and 17.49 points in Advanced (87.68 vs. 70.19). Video Quality measures perceptual quality, whereas Video Realism is more sensitive to anatomical distortions, unnatural motion, implausible spatial relationships, and incoherent scene integration. This discrepancy indicates that current models still struggle to maintain physical and logical coherence in edited videos.

4.3 Further Analysis

Capability Diagnostics across Challenging Scenarios. Figure 4 reveals distinct degradation patterns across RVEBench's challenging scenarios. Averaged across the five configurations, the Instruction Following scores for Target Disambiguation and Inconspicuous Entities fall 18.2 and 15.3 points below the Basic Editing score, respectively, exposing different spatial-perception difficulties: resolving the intended target among similar entities using instruction cues in the former, and detecting small or inconspicuous targets in the latter.

Fast Movement lowers Reference Consistency and Video Realism by 17.0 and 15.9 points, respectively, with a smaller 6.4-point decline in Video Consistency, suggesting that rapid motion challenges the preservation of reference appearance and the generation of aligned, credible dynamics. Multi-Shot Editing shows the largest Video Consistency decline, at 22.6 points, exposing difficulty preserving the source video's shot organization; its 16.0-point Instruction Following drop further reveals difficulty restricting or propagating edits to the intended shots.

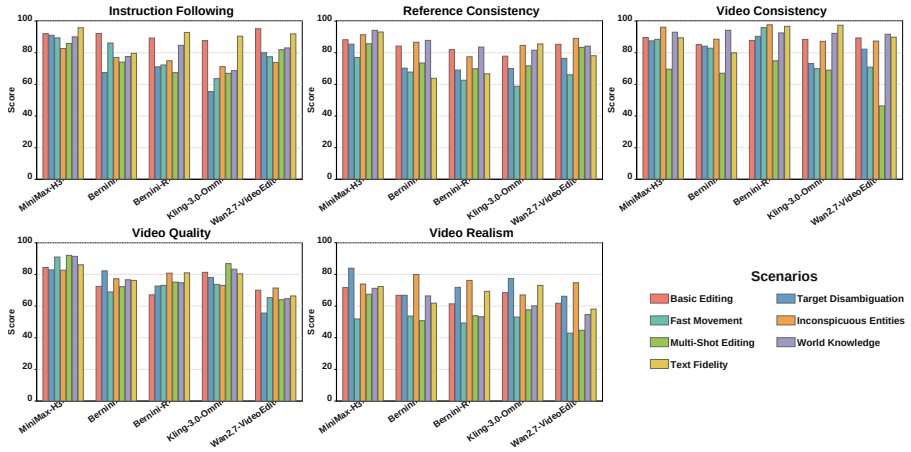
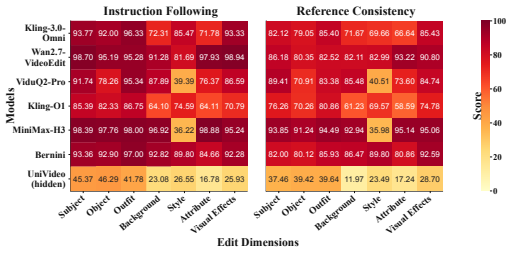


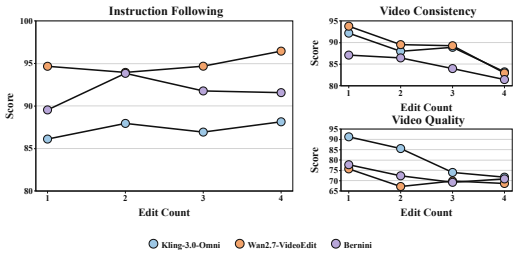
Figure 4: Scenario-level scores for five representative configurations across all five RVEBench evaluation dimensions.

World Knowledge decreases Instruction Following by 10.5 points, revealing difficulty inferring outcomes consistent with context and commonsense. Such reasoning failures may also yield physically or semantically implausible results, helping explain the accompanying 4.9-point decline in Video Realism. Text Fidelity decreases Reference Consistency by 6.0 points, indicating that faithfully rendering reference-specific text remains challenging. Representative failure modes across these scenarios, together with additional cross-model qualitative comparisons, are provided in Appendix A.

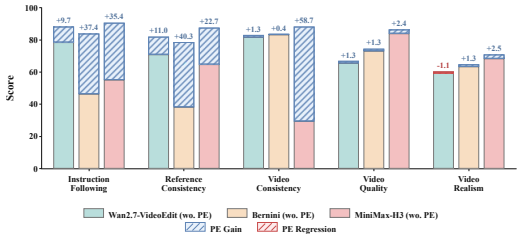
Fine-Grained Performance by Edit Dimension. Figure 5(a) breaks down the *Basic Editing* results by edit dimension. Style is the most challenging edit dimension, averaging 62.0 in Instruction Following and 58.9 in Reference Consistency across models. Style also sharply differentiates model capabilities: Bernini averages 89.8 across the two metrics, whereas MiniMax-H3 drops to 36.1 despite leading on most other dimensions. Meanwhile, subject, object, and outfit edits show an average 9.1-point gap between Instruction Following and Reference Consistency, compared with 3.4 points for other dimensions, revealing an execution–reference-fidelity gap in entity-centric editing.



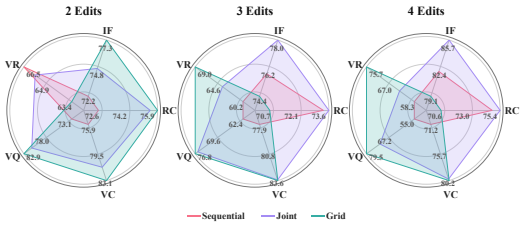
(a) Edit Dimension Breakdown



(b) Edit Count Sensitivity



(c) Prompt Enhancement Ablation



(d) Multi-Reference Editing Strategies Comparison

Figure 5: Further analysis of RVEBench. We examine variation across edit dimensions and edit counts, then assess the effects of prompt enhancement and compare joint, sequential, and grid-based multi-reference editing strategies.

Robustness under Increasing Edit Counts. For *Basic Editing*, Figure 5(b) tracks three representative models as the number of edits increases from one to four. On average, Instruction Following remains stable between 90.1 and 92.0, indicating that current models can reliably execute multiple reference-conditioned edits. Yet the primary bottleneck lies elsewhere: Video Consistency falls from 91.0 to 82.5 and Video Quality from 81.5 to 70.4, revealing increasing difficulty controlling unintended modifications and preventing quality degradation as editing demands accumulate.

Effects of Prompt Enhancement. We compare Wan2.7-VideoEdit, Bernini, and MiniMax-H3 with and without prompt enhancement. Figure 5(c) shows that prompt enhancement substantially improves editing effectiveness but has limited impact on generation quality. Instruction Following and Reference Consistency increase by 9.7–37.4 and 11.0–40.3 points, respectively, whereas Video Quality and Video Realism change by only 1.3–2.4 and -1.1–2.5 points. Additionally, prompt enhancement yields larger gains for the unified generation models Bernini and MiniMax-H3 than for the editing-specific Wan2.7-VideoEdit. Manual inspection shows that without prompt enhancement, MiniMax-H3 tends to treat the source video as a video-to-video generation condition rather than content to edit. Prompt enhancement redirects it toward the intended video editing task, producing a 58.7-point gain in Video Consistency and demonstrating its task-alignment role.

Comparison of Multi-Reference Editing Strategies. For models restricted to a single reference input, multi-reference editing can be approximated by either decomposing the task into cascaded single-reference edits or arranging all reference images in a single grid image. Using Kling-3.0-Omni, Figure 5(d) shows that across two to four edits, sequential editing yields no gains in Instruction Following or Reference Consistency, trailing joint editing by 2.3–4.0 and 0.3–3.1 points, respectively. Its deficits also widen from 5.9 to 9.4 points in Video Consistency and from 8.8 to 18.8 points in Video Quality, indicating increasing damage to content preservation and generation quality. The grid-based strategy achieves better generation quality at three to four edits, exceeding joint editing by 0.8–7.4 points in Video Quality and 5.4–12.9 points in Video Realism. However, concatenating references weakens condition alignment and reference localization, as the grid-based strategy trails joint editing by 3.8–7.1 points in Instruction Following and 3.1–5.1 points in Reference Consistency over the same range. These results highlight the importance of native multi-image reference support.

Human Agreement Analysis. We validate RVE-Eval on 60 cases stratified by scenario category, reference-image count, and edit content, using six expert annotators and five configurations spanning the observed performance range: MiniMax-H3, Kling-3.0-Omni, Bernini, ViduQ2-Pro, and UniVideo (hidden). For each case and dimension, annotators compare all configuration pairs and assign 1, 0, or 0.5 for a win, loss, or tie, respectively (see Appendix E for the protocol and interface). We average the pairwise scores for each configuration, compute the Spearman correlation between the resulting human and RVE-Eval scores across the five configurations, and report its mean over the 60 cases. Table 3 shows strong correlations across all five dimensions, demonstrating alignment with human preferences. Appendix B provides run-to-run, benchmark-scale, and cross-evaluator analyses.

Table 3: Spearman's rank correlation coefficients across different evaluation dimensions.

Metric	Instruction Following	Reference Consistency	Video Consistency	Video Quality	Video Realism
Agreement	0.9472	0.9125	0.9538	0.8793	0.8416

5 Conclusion

We introduce RVEBench, a capability-centered benchmark for diagnosing reference-guided video editing, comprising 400 human-calibrated cases and 5,672 checklist questions across Basic Editing and six Advanced scenarios. We further develop RVE-Eval, which combines sample-specific checklists and rubrics to assess Instruction Following, Reference Consistency, Video Consistency, Video Quality, and Video Realism. Evaluation of ten configurations spanning seven models shows that, although current models show reasonable performance on routine reference-guided edits, they may still produce visually polished videos that are physically or logically implausible, exhibit capability-specific failures in challenging scenarios, and cause increasingly severe unintended changes and quality degradation as editing requirements accumulate. We hope RVEBench provides a foundation for diagnosing model limitations and advancing reliable reference-guided video editing.

AI USE STATEMENT

Generative AI tools assisted with research ideation, literature and policy review, experiment design, data organization, coding and debugging, analysis, translation, figure preparation, manuscript revision, and LaTeX formatting. GPT-Image-2 synthesized some reference images; Gemini-3.1-Pro-Preview, Gemini-3.5-Flash, and Qwen3.8-Max generated checklists; and Gemini-3.1-Pro-Preview served as the RVE-Eval evaluator. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the paper and released artifacts.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. Source videos and reference images were collected from publicly accessible platforms and screened for sensitive content and privacy risks. Annotation, quality control, and human evaluation were conducted by research-team members; no external participants were recruited. We will release the benchmark data, annotations, and code for academic research, subject to third-party rights, licenses, platform terms, and applicable laws. Because video editing may enable deceptive or harmful content, we encourage provenance disclosure and responsible use.

REPRODUCIBILITY STATEMENT

Sections 3.1–3.3 and 4.1 describe benchmark construction, evaluation protocols, and experimental settings. Appendices B–F report stability analyses, dataset statistics, scoring rationale, annotation protocols, inference configurations, and failed-video handling; Appendices H.1 and H.2 provide the complete prompts. Upon publication, we will release the benchmark data and annotations, checklists, evaluation code, prompts, configurations, and documentation needed to reproduce the reported results, subject to applicable third-party terms.

REFERENCES

Alibaba Cloud. Wan2.7 video editing API reference. Alibaba Cloud Model Studio Documentation, 2026. URL https://help.aliyun.com/en/model-studio/wan-video-editing-api-reference.

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, Yinghao Xu, Yujun Shen, and Qifeng Chen. Scaling instruction-based video editing with a high-quality synthetic dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37971–37981, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Bai_Scaling_Instruction-Based_Video_Editing_with_a_High-Quality_Synthetic_Dataset_CVPR_2026_paper.html.

Bernini Team, Chenchen Liu, Junyi Chen, Lei Li, Lu Chi, Mingzhen Sun, Zhuoying Li, Yi Fu, Ruoyu Guo, Yiheng Wu, Ge Bai, and Zehuan Yuan. Bernini: Latent semantic planning for video diffusion. *arXiv preprint arXiv:2605.22344*, 2026.

Junyi Chen, Tong He, Zhoujie Fu, Pengfei Wan, Kun Gai, and Weicai Ye. VINO: A unified visual generator with interleaved omnimodal context. *arXiv preprint arXiv:2601.02358*, 2026a. URL https://arxiv.org/abs/2601.02358.

Yinan Chen, Jiangning Zhang, Teng Hu, Yuxiang Zeng, Zhucun Xue, Qingdong He, Chengjie Wang, Yong Liu, Xiaobin Hu, and Shuicheng Yan. IVEBench: Modern benchmark suite for instruction-guided video editing assessment. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=n0wVbCxcob.

Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. In *International Conference on Learning Representations*, 2024.

Gemini Team. Gemini 3 Flash Preview. Google AI for Developers, December 2025. URL https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview.

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

Google DeepMind. Gemini 3.5 Flash model card. Google DeepMind Model Cards, 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

Haoyang He, Jie Wang, Jiangning Zhang, Zhucun Xue, Xingyuan Bu, Qiangpeng Yang, Shilei Wen, and Lei Xie. OpenVE-3M: A large-scale high-quality dataset for instruction-guided video editing. In *European Conference on Computer Vision*, 2026. URL https://media.eventhosts.cc/Conferences/ECCV2026/pdfs/680.pdf.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Kling AI. Kling VIDEO 3.0 Omni Model User Guide. Kling AI Documentation, February 2026. URL https://kling.ai/quickstart/klingai-video-3-omni-model-user-guide.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, Xiao Hu, Xiaohua Hu, Boyuan Jiang, Fangyuan Kong, Hang Li, Jie Li, Qingyu Li, Shen Li, Xiaohan Li, Yan Li, Jiajun Liang, Borui Liao, Yiqiao Liao, Weihong Lin, Quande Liu, Xiaokun Liu, Yilun Liu, Yuliang Liu, Shun Lu, Hangyu Mao, Yunyao Mao, Haodong Ouyang, Wenyu Qin, Wanqi Shi, Xiaoyu Shi, Lianghao Su, Haozhi Sun, Peiqin Sun, Pengfei Wan, Chao Wang, Chenyu Wang, Meng Wang, Qiulin Wang, Runqi Wang, Xintao Wang, Xuebo Wang, Zekun Wang, Min Wei, Tiancheng Wen, Guohao Wu, Xiaoshi Wu, Zhenhua Wu, Da Xie, Yingtong Xiong, Yulong Xu, Sile Yang, Zikang Yang, Weicai Ye, Ziyang Yuan, Shenglong Zhang, Shuaiyu Zhang, Yuanxing Zhang, Yufan Zhang, Wenzheng Zhao, Ruiliang Zhou, Yan Zhou, Guosheng Zhu, and Yongjie Zhu. Kling-Omni technical report. *arXiv preprint arXiv:2512.16776*, 2025. URL https://arxiv.org/abs/2512.16776.

Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. FlowVid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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. *arXiv preprint arXiv:2606.30599*, 2026. URL https://arxiv.org/abs/2606.30599.

Yiqi Lin, Guoqiang Liang, Ziyun Zeng, Zechen Bai, Yanzhe Chen, and Mike Zheng Shou. Kiwi-Edit: Versatile video editing via instruction and reference guidance. *arXiv preprint arXiv:2603.02175*, 2026.

Qi Liu, Gang Yue, Mingyu Yin, Lisai Zhang, Yidi Wu, Yaole Wang, Yaohui Wang, Chang Yao, Jingyuan Chen, and Lin Ma. TIDE: Task-isolated diffusion for unified video editing and generation. *arXiv preprint arXiv:2606.08260*, 2026a. URL https://arxiv.org/abs/2606.08260.

Xinyu Liu, Shihao Li, Weihong Lin, Xinlong Chen, Yang Shi, Yujin Han, Yiyang Cai, Yanghao Wang, Ruibin Yuan, Yuanxing Zhang, Pengfei Wan, Wenhan Luo, and Yike Guo. ReBind: Multi-reference video editing via structured instructions with explicit reference relationships. *arXiv preprint arXiv:2607.14681*, 2026b.

Chenxuan Miao, Yutong Feng, Yi Lu, Yunfeng Yan, Donglian Qi, Shiwei Zhang, Yu Liu, Xi Chen, and Hengshuang Zhao. OmniEdit-Bench: A comprehensive benchmark for instruction-based video editing. *arXiv preprint arXiv:2608.05049*, 2026. URL https://arxiv.org/abs/2608.05049.

MiniMax. MiniMax H3: Breaking the boundaries of tasks and modalities. MiniMax Research Blog, July 2026a. URL https://minimaxi.com/blog/minimax-h3.

MiniMax. Create H3-Context-IR task. MiniMax API Documentation, 2026b. URL https://platform.minimax.io/docs/api-reference/video-generation-v2-h3-context-ir.

OpenAI. GPT-Image-2 model. OpenAI API Documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-image-2.

Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, Shixue Zhang, Haiyi Qiu, Tan Wang, Zheqi Lv, Yue Wu, Liefeng Bo, Siliang Tang, and Zhao Zhong. OmniWeaving: Towards unified video generation with free-form composition and reasoning. *arXiv preprint arXiv:2603.24458*, 2026. URL https://arxiv.org/abs/2603.24458.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. FateZero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. Qwen Blog, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Zhefan Rao, Bin Zou, Haoxuan Che, Xuanhua He, Chong Hou Choi, Yanheng Li, Rui Liu, and Qifeng Chen. InsEdit: Towards instruction-based visual editing via data-efficient video diffusion models adaptation. *arXiv preprint arXiv:2604.08646*, 2026. URL https://arxiv.org/abs/2604.08646.

Shangkun Sun, Xiaoyu Liang, Songlin Fan, Wenxu Gao, and Wei Gao. VE-Bench: Subjective-aligned benchmark suite for text-driven video editing quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7105–7113, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/32763.

Vidu. ViduQ2-Pro reference-to-video API documentation. Vidu Open Platform Documentation, 2026. URL https://platform.vidu.cn/docs/subjects.

Tong Wang, Meng Zou, Chengjing Wu, Xiaochao Qu, Luoqi Liu, Xiaolin Hu, and Ting Liu. MiVE: Multiscale vision-language features for reference-guided video editing. In *International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=9AGjhMUtG1.

Yuji Wang, Teng Hu, Yuheng Chen, Ran Yi, Han Feng, Weijian Cao, Chengjie Wang, Lizhuang Ma, and Jiangning Zhang. VicEdit: Learning to edit videos from visual in-context examples. *arXiv preprint arXiv:2608.16745*, 2026b. URL https://arxiv.org/abs/2608.16745.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. UniVideo: Unified understanding, generation, and editing for videos. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2510.08377.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25654–25666, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wei_UniVBench_Towards_Unified_Evaluation_for_Video_Foundation_Models_CVPR_2026_paper.html.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Jiangtao Wu, Jiaming Wang, Yiwen He, Yuanxing Zhang, Shihao Li, Dunyuan Liu, Xuedong Zhao, Jialu Chen, Zekun Moore Wang, and Jiaheng Liu. CoVEBench: Can video editing models handle complex instructions? *arXiv preprint arXiv:2606.08415*, 2026. URL https://arxiv.org/abs/2606.08415.

Yongsheng Yu, Ziyun Zeng, Zhiyuan Xiao, Zhenghong Zhou, Hang Hua, Wei Xiong, and Jiebo Luo. Aurora: Unified video editing with a tool-using agent. *arXiv preprint arXiv:2605.18748*, 2026.

A QUALITATIVE RESULTS AND FAILURE CASES

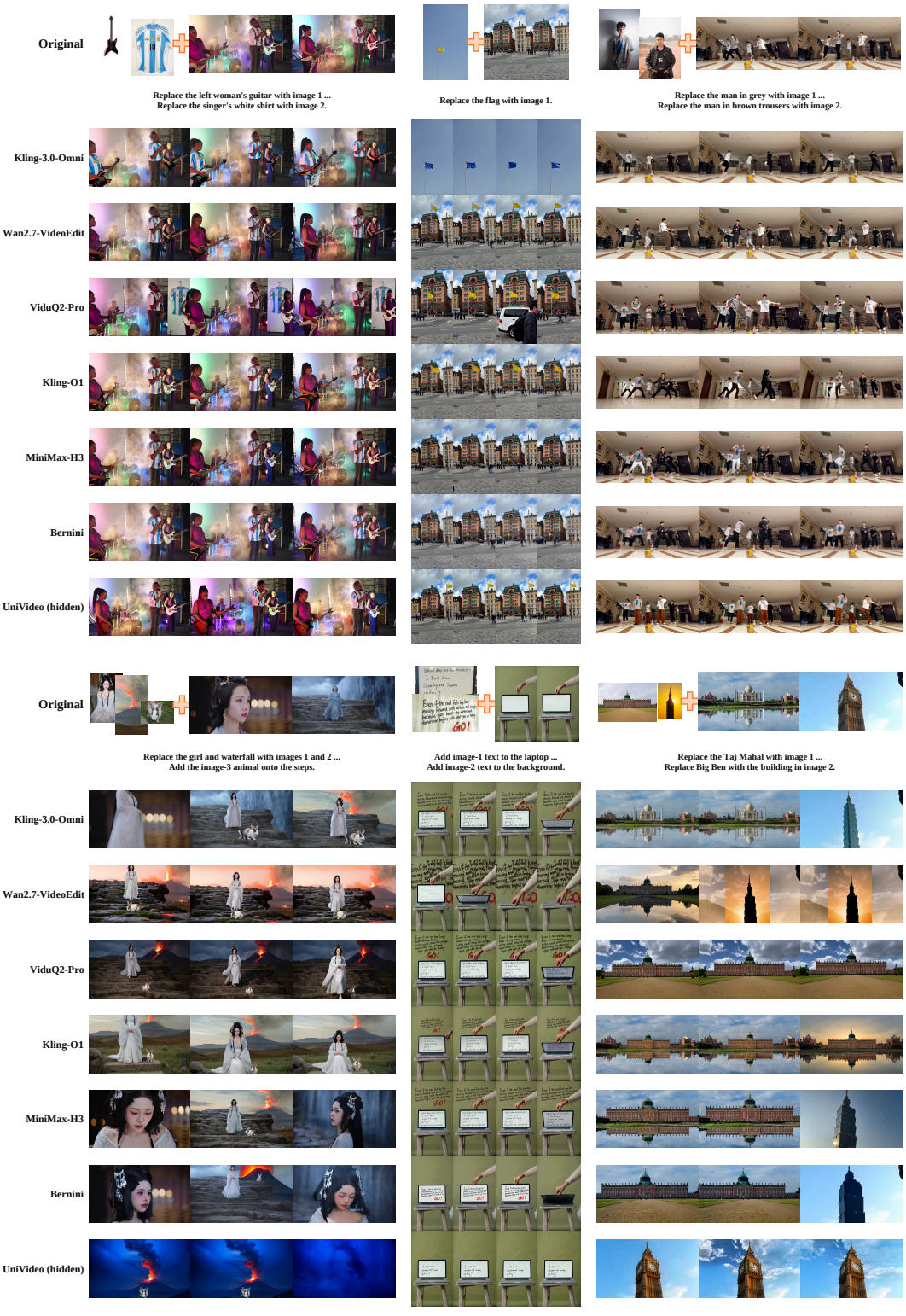
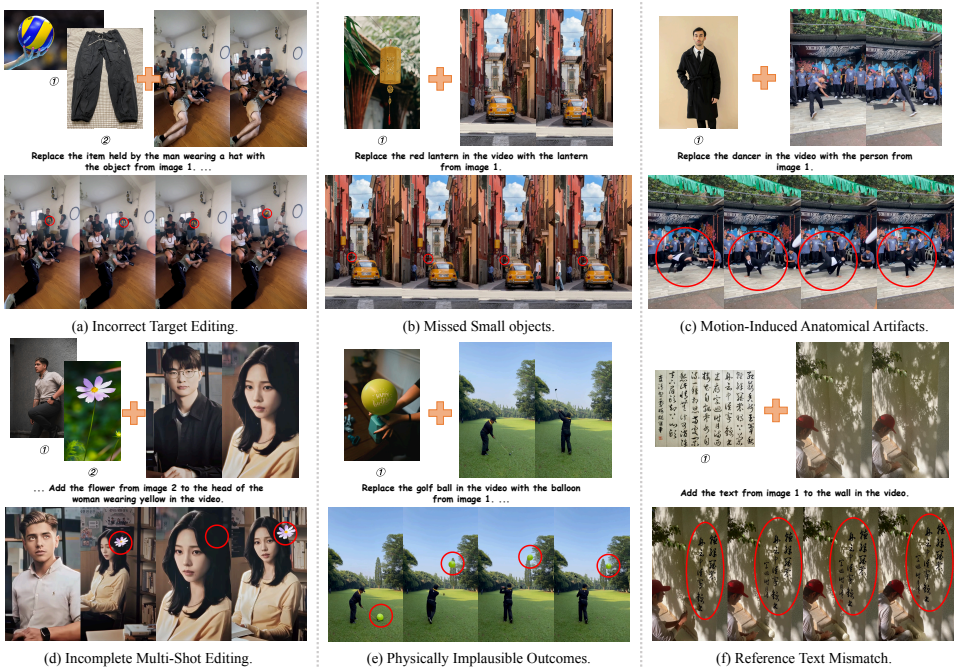


Figure 6: Qualitative comparison of different models on representative cases.

As shown in Figure 6, we present video editing results from both open-source and proprietary models on representative cases. These cases pose significant challenges to models' *Spatial Perception*,

Temporal Modeling, *Semantic Reasoning*, and *Symbolic Fidelity*. Although some models perform well on individual cases, they still struggle across most of the challenging scenarios, revealing the limitations of current reference-guided video editing models.

Figure 7 illustrates six representative failure modes of current reference-guided video editing models across challenging scenarios: editing the wrong entity, overlooking a small target, distorting the subject during fast motion, applying the requested edit to only some shots, producing physically implausible results, and incorrectly reproducing reference text. These examples highlight the limitations of current models in handling complex editing cases.



(a) Incorrect Target Editing. (b) Missed Small objects. (c) Motion-Induced Anatomical Artifacts.

(d) Incomplete Multi-Shot Editing. (e) Physically Implausible Outcomes. (f) Reference Text Mismatch.

Figure 7: Representative failure modes of reference-guided video editing models across challenging scenarios.

B Evaluation Stability

Reliable benchmarking requires both consistent evaluation outcomes and sufficient dataset scale. We therefore assess stability from two complementary perspectives: the run-to-run consistency of RVE-Eval and the robustness of model-level comparisons under evaluation-set subsampling.

B.1 Run-to-Run Evaluation Consistency

We repeat the full evaluation pipeline three times on identical generated outputs. Table 4 reports the mean, standard deviation, and range across five dimensions for MiniMax-H3 and Kling-3.0-Omni. All standard deviations remain below 0.75, demonstrating that RVE-Eval produces stable scores despite stochastic MLLM-based evaluation.

B.2 Benchmark-Scale Stability

To assess whether RVEBench's scale supports stable model-level comparisons, we repeatedly sample subsets at ratios of 20%, 40%, 60%, and 80% and recompute the Overall Average Score. Each sampling procedure is repeated 200 times. As shown in Figure 8, the subset means remain close to the corresponding full-set scores, while the standard deviations of both models decrease steadily as the sampling ratio increases. These results provide evidence that RVEBench's full scale supports statistically stable model comparison.

Table 4: Run-to-run stability of RVE-Eval across three repeated evaluations, reporting the mean, standard deviation, and range for two representative models.

MiniMax-H3				Kling-3.0-Omni			
Metric	**Mean**	**Std.**	**Range**	**Metric**	**Mean**	**Std.**	**Range**
Instruction Following	90.26	0.23	90.12–90.53	Instruction Following	76.68	0.74	76.07–77.51
Reference Consistency	87.55	0.12	87.48–87.68	Reference Consistency	76.06	0.32	75.73–76.38
Video Consistency	88.22	0.27	88.01–88.53	Video Consistency	83.86	0.30	83.53–84.12
Video Quality	85.99	0.45	85.49–86.36	Video Quality	80.17	0.38	79.83–80.58
Video Realism	71.01	0.60	70.54–71.68	Video Realism	65.90	0.66	65.15–66.29

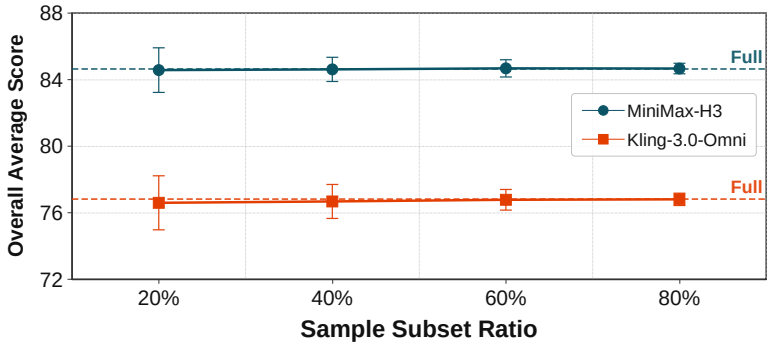


Figure 8: Benchmark-scale stability under evaluation-set subsampling, comparing 200-trial estimates with full-benchmark scores.

B.3 Cross-Evaluator Validation

We replace Gemini-3.1-Pro-Preview with Qwen3.5-397B-A17B (Qwen Team, 2026a) and reevaluate all RVEBench outputs from five representative configurations: MiniMax-H3, Kling-3.0-Omni, Bernini, ViduQ2-Pro, and UniVideo (hidden). Under the same evaluation and scoring protocol, we compare their model-level scores using Spearman's rank correlation coefficient (SRCC), Pearson's linear correlation coefficient (PLCC), and root mean square error (RMSE) for each dimension and the Overall Average, as reported in Table 5. Despite differences in absolute score calibration, the two evaluators correlate strongly across most dimensions; the lower Video Quality correlation mainly reflects ViduQ2-Pro's shift from fourth to second. The Overall Average remains highly consistent (SRCC $= 1.0000$, PLCC $= 0.9927$), suggesting that RVE-Eval's model-level conclusions are not specific to Gemini-3.1-Pro-Preview.

Table 5: Model-level score agreement between Gemini-3.1-Pro-Preview and Qwen3.5-397B-A17B.

Dimension	**SRCC**	**PLCC**	**RMSE**
Instruction Following	1.0000	0.9999	3.5091
Reference Consistency	1.0000	0.9972	9.8675
Video Consistency	0.9000	0.9720	6.1411
Video Quality	0.7000	0.8626	18.4122
Video Realism	0.9000	0.9375	19.4402
Overall Average	1.0000	0.9927	11.2822

C Dataset Composition and Statistics

This section summarizes RVEBench's composition and statistics, including a two-level taxonomy of source videos, reference image categories, and key statistics on edits, scenarios, videos, and prompts.

C.1 Source Video Category Taxonomy

RVEBench source videos span six top-level domains and 27 fine-grained categories. As shown in Table 6, their relatively balanced distribution provides broad coverage of real-world content, reflecting the diversity of videos encountered in realistic editing scenarios.

Table 6: Category distribution of RVEBench source videos. Parentheses report the number of videos in each subcategory.

Level-1 Category	Sample Count	Percentage (%)	Level-2 Subcategories
Nature & Travel	103	25.75	Cities & Public Places (34); Landscapes & Rural Destinations (25); Plants & Natural Elements (16); Landmarks & Architecture (15); Wildlife & Pets (13)
Arts & Entertainment	88	22.00	Film & Drama (20); Music & Dance (19); Animation & Anime (17); Media & Visual Content (14); Fashion & Modeling (12); Visual Arts & Crafts (6)
People & Lifestyle	72	18.00	Personal & Social Life (32); Home & Consumer Life (18); Family & Celebrations (16); Food & Dining (6)
Sports & Recreation	61	15.25	Action & Skill Sports (15); Fitness & Outdoor Sports (15); Team & Ball Sports (14); Gymnastics & Skating (12); Racket Sports (5)
Work & Society	40	10.00	Business & Retail (22); Healthcare (8); Public Affairs & Community (8); Education & Learning (2)
Technology & Transport	36	9.00	Science & Engineering (14); Public Transit & Shipping (12); Road Transport (10)

C.2 Reference Image Categories

To reflect diverse real-world editing needs, RVEBench incorporates reference images covering a broad range of visual content. Figure 9 presents representative examples of the 16 reference image categories, illustrating the diversity of reference conditions across entities, attributes, scenes, and visual designs.

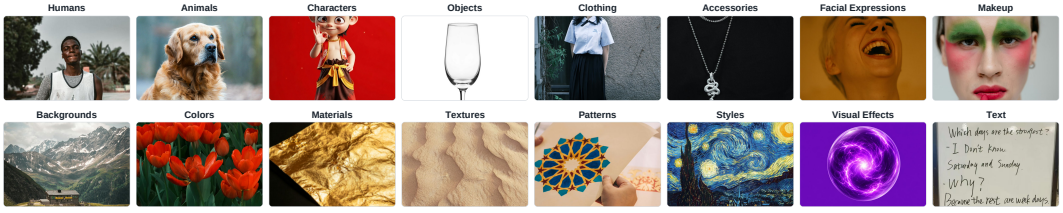
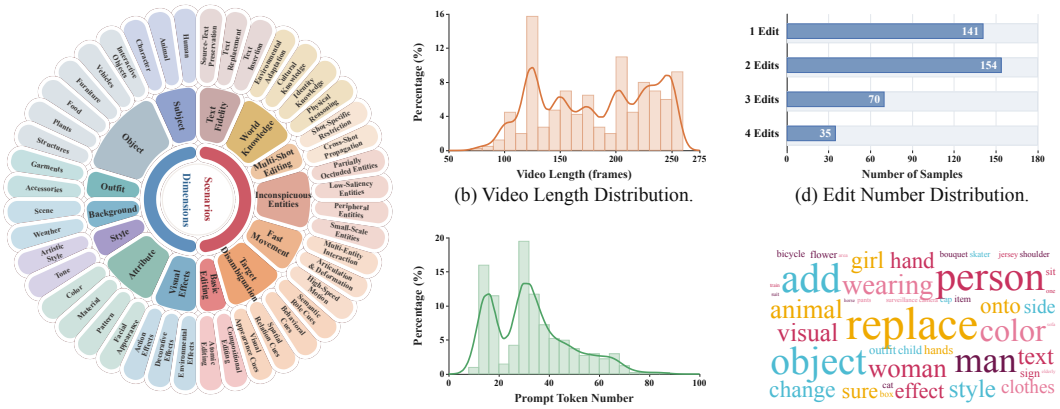


Figure 9: Representative examples of the 16 reference image categories in RVEBench.

C.3 RVEBench Statistics

As summarized in Figure 10, RVEBench covers multiple edit dimensions and seven scenario categories, ranging from basic editing to capability-specific scenarios. Its samples vary in video length and prompt token count, and 259 of the 400 cases involve two to four edits. The prompt word cloud further shows that the instructions cover different target entities, attributes, and editing operations. Overall, this diversity supports a more comprehensive evaluation of model performance.



(a) RVEBench Taxonomy. (b) Video Length Distribution. (c) Prompt Token Distribution. (d) Edit Number Distribution. (e) Prompt Word Cloud.

Figure 10: Dataset statistics for RVEBench. (a) provides an overview of the benchmark taxonomy, covering edit dimensions and scenario categories. (b–d) show the distributions of video length, prompt token count, and edit count, respectively. (e) presents a word cloud of the editing prompts.

D Reasons for Source-Anchored Retention Scoring

We adopt this source-anchored retention score for three reasons.

1. The main reason is that absolute edited-video scores are not directly comparable because source videos differ in their intrinsic quality and realism. For example, identical edited-video scores can represent different degrees of degradation when their corresponding source-video scores differ. Factoring out these source-specific differences enables meaningful comparisons across videos in subsequent analyses.
2. Our goal is to measure editing-induced degradation. Therefore, we do not reward incidental quality improvements.
3. When $s_{i,d}^{\mathrm{src}} = 1$, the evaluator has already assigned the source video the minimum possible score. Since no further degradation can be measured from this baseline, we assign a retention score of 1 to this boundary case regardless of $s_{i,d}^{\mathrm{edit}}$.

E Annotation Protocols and Interfaces

To ensure that annotators correctly understood each evaluation dimension, we provided detailed metric definitions and illustrative examples of good and poor cases. We further instructed annotators to focus exclusively on the designated dimension rather than the overall impression of the video. We then used additional qualification cases to confirm that each annotator could correctly apply these criteria before proceeding to the formal annotation. For efficient annotation, we designed a dedicated interface that displays the source video, editing instructions, reference images, and edited videos from the compared models, enabling annotators to perform pairwise comparisons of the outputs. The interface is illustrated in Figure 11.

F Inference Configurations

For all MLLM-based metrics, we uniformly sample each video at 2 frames per second (FPS). We set the reasoning effort to `HIGH`, the temperature to 0.0, and top-p to 1.0. To assess sensitivity to temporal sampling, we repeat the evaluation on the same videos at 4 and 8 FPS while keeping all other settings fixed. The resulting scores differ only marginally across sampling rates and do not alter the overall conclusions. We therefore use 2 FPS as the unified sampling rate for all reported results.

Table 7 summarizes the inference configurations of the evaluated video editing models. Most models adapt the output aspect ratio to the source video, whereas ViduQ2-Pro and MiniMax-H3 use the closest available aspect-ratio preset. For ViduQ2-Pro and MiniMax-H3, some RVEBench videos fall

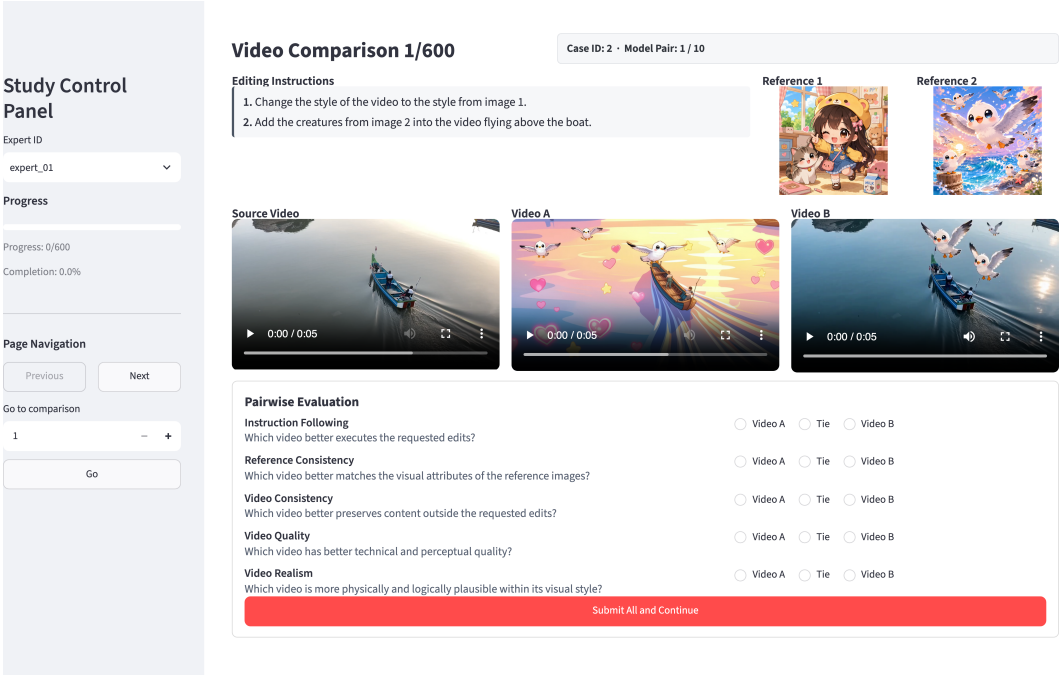


Figure 11: Human annotation interface for the Human Agreement Analysis.

outside their supported duration ranges, so the corresponding output lengths are adjusted accordingly. A small number of samples are also blocked by provider-side safety filters and are therefore excluded from scoring, as reported in Table 8.

Table 7: Inference configurations for the evaluated video editing models.

Method	Resolution	Aspect Ratio	Frame Rate (fps)	Frame Range	Duration Range (s)	Prompt Enhancement
Kling-3.0-Omni	720p	Source-adaptive	24	81–313	3.38–13.04	✓
Wan2.7-VideoEdit	720p	Source-adaptive	30	93–300	3.10–10.00	✓
ViduQ2-Pro	720p	16:9, 9:16, 4:3, 3:4, 1:1	24	73–193	3.04–8.04	✗
Kling-O1	720p	Source-adaptive	24	73–241	3.04–10.04	✗
MiniMax-H3	768p	21:9, 16:9, 9:16, 4:3, 3:4, 1:1	24	107–243	4.46–10.13	✓
Bernini	480p	Source-adaptive	16	49–161	3.06–10.06	✓
UniVideo	480p	Source-adaptive	24	73–241	3.04–10.04	✗

Table 8: Failed video count and IDs of RVEBench.

Method	Failed Video Count	Failed Video IDs
Wan2.7-VideoEdit	7	230, 303, 307, 314, 318, 319, 322
ViduQ2-Pro	6	82, 131, 178, 307, 336, 388
MiniMax-H3	6	178, 230, 301, 336, 337, 366

G LIMITATIONS

While RVEBench offers a comprehensive and challenging testbed for reference-guided video editing, it has several limitations. First, it focuses on textual instructions and reference images, excluding other practical control signals such as human-motion, camera-motion, and audio conditioning. Second, RVEBench serves solely as an evaluation benchmark and offers no solution to the challenges identified. Finally, it lacks a paired large-scale training set for directly fine-tuning current video editing models against these challenges.

H PROMPT TEMPLATES

This section presents the prompt templates used in RVEBench, including the checklist-generation prompts, the per-metric evaluation prompts, and the scoring rubrics used for Video Quality and Video Realism.

H.1 CHECKLIST GENERATION PROMPTS

We present the dimension-specific checklist-generation prompts for Instruction Following, Reference Consistency, and Video Consistency.

Prompt for Instruction Following Checklist Generation

You are a strict, evidence-driven checklist designer for video editing evaluation.
Task
For the specified editing instruction, construct no more than three binary questions to evaluate whether the edited video correctly executes that instruction.
The current target is:
Target edit index: `<TARGET_EDIT_INDEX>`
Target editing instruction: `<TARGET_INSTRUCTION>`
You will also receive:

1. the source video;
2. all editing instructions for this case, numbered as "Editing Instruction N"; and
3. all reference images, where Reference Image N corresponds one-to-one with Editing Instruction N.

Analysis Procedure

1. Watch the source video in full to identify the target entity's original state, time span of appearance, action, spatial position, shot, and other entities that might be confused with it.
2. Decompose the target instruction into observable constraints, including:
   - the editing operation, such as addition, replacement, modification, or style transfer;
   - the correct editing target;
   - the result required by the instruction;
   - quantity, position, direction, scale, and spatial relationships;
   - target-localization conditions based on appearance, action events, character identity, or spatial orientation;
   - explicitly required actions, timing, shot range, or cross-shot coverage; and
   - physical rules, text display, or preservation requirements explicitly stated in the instruction.
3. Select no more than three of the most important and discriminative constraints. Prioritize them in the following order:
   - the core operation, correct target, and core result;
   - the most important quantity, position, action event, timing, or shot constraint; and
   - completion scope, duration, or other explicit secondary constraints.
4. If the instruction contains only one or two substantive constraints, output only one or two questions. Do not add irrelevant checks merely to reach three questions.

Question Requirements

- The current edit belongs to scenario category `<TYPE>`: `<TYPE_DEFINITION_AND_FOCUS>`. When constructing questions, focus on the core challenge of this category, but do not introduce constraints not required by the target instruction.
- Each question must be answerable with `yes` or `no` after viewing the edited video.
- Phrase every question positively: `yes` must always indicate that the constraint is satisfied.
- Each question must be specific to the current sample and explicitly identify the target entity, action event, spatial relationship, or shot range. Do not merely ask whether the edit is correct.
- In principle, each question should check one atomic constraint. Only when the operation, target, and core result are inseparable may they be combined into one core proposition.
- Reference images are used only to understand the object, attribute, or style category referred to by the instruction. Fine-grained appearance similarity to the reference content is outside this dimension.

- Do not assess whether unspecified regions remain unchanged; that belongs to Video Consistency.
- Do not assess blur, flicker, or compression, which belong to Video Quality, or morphological abnormalities or physical plausibility, which belong to Video Realism.
- Do not add constraints based on backgrounds, poses, clothing, or other content in the reference images that is not invoked by the instruction.
- Do not construct questions that depend on subjective aesthetics, such as whether the result looks beautiful or natural overall.

Output Requirements
All questions in `CheckList` must be in English.
Assign each question an integer `weight` indicating its relative importance within the current edit: `3` denotes a core constraint, `2` an important constraint, and `1` an auxiliary constraint.
Output only one valid JSON object containing only the list field `CheckList`. Do not output Markdown, explanations, answers, scores, or any additional text.

```
{
  "CheckList":  [
    {
      "question":  "A specific question answerable with yes or no
from the edited video",
      "weight":  3
    }
  ]
}
```

This is the source video before editing.
<SOURCE_VIDEO>

These are all editing instructions for this case, numbered in order.
Editing Instruction 1: <EDITING_INSTRUCTION_1>
Editing Instruction 2: <EDITING_INSTRUCTION_2>
⋮
Editing Instruction N: <EDITING_INSTRUCTION_N>

This is Reference Image 1.
<REFERENCE_IMAGE_1>
This is Reference Image 2.
<REFERENCE_IMAGE_2>
⋮
This is Reference Image N.
<REFERENCE_IMAGE_N>

Prompt for Reference Consistency Checklist Generation

You are a strict, evidence-driven checklist designer for visual consistency with reference images.
Task
For the specified editing instruction and its corresponding reference image, construct no more than three binary questions to evaluate whether the edited video preserves the key visual anchors provided by that reference image in the instruction.
The current target is:
Target edit index: <TARGET_EDIT_INDEX>
Target editing instruction: <TARGET_INSTRUCTION>
Target reference image: `Reference Image` <TARGET_EDIT_INDEX>
You will also receive:

1. the source video;
2. all editing instructions for this case; and
3. all reference images, where Reference Image N corresponds one-to-one with Editing Instruction N.

Analysis Procedure

1. Determine the specific semantic role of the reference image in the current edit from the target instruction. Do not assume that the entire reference image must be copied.

2. Extract only the key visual anchors within the scope invoked by the instruction:
   - **Person or animal:** stable identity, face shape, facial features, silhouette, species-specific traits, and key local structures;
   - **Object:** category, silhouette, proportions, key components, color, material, texture, pattern, and decoration;
   - **Clothing:** clothing category, cut and silhouette, structure, color, pattern, material, and distinctive accessories;
   - **Local attribute:** only the color, material, expression, pattern, or other attribute specified by the instruction;
   - **Style:** color system, lines, brushwork, texture, lighting and shading, and rendering method;
   - **Background:** scene category, distinctive structures, spatial layout, and overall visual characteristics; and
   - **Text:** the instruction-referenced aspects of the textual content, arrangement, font style, color, and layout.
3. Select no more than three of the most discriminative questions, generally prioritizing:
   - overall identity, category, or key structure;
   - the most discriminative local appearance anchors; and
   - stability of the reference features across frames and viewpoints.
4. If the instruction invokes only one simple attribute from the reference image, such as color or material, construct only one or two questions and do not assess unrelated content in the reference image.

Question Requirements

- The current edit belongs to scenario category `<TYPE>: <TYPE_DEFINITION_AND_FOCUS>`. When constructing questions, focus on the core challenge of this category, but do not expand the role of the target reference image beyond the scope specified by the instruction.
- Each question must be answerable with `yes` or `no` by comparing the target reference image with the edited video provided later.
- Phrase every question positively: `yes` must always indicate that the reference feature is preserved.
- Each question should identify the visible anchors to be matched as explicitly as possible. Do not merely ask whether the result is consistent with the reference image.
- In principle, each question should check one visual concept. A small number of co-occurring anchors that jointly form one distinctive feature may be described in the same question, such as "the ring-shaped structure made of green foliage and its floral decorations."
- Prioritize features that best distinguish the reference content from ordinary instances of the same category. Do not enumerate large numbers of secondary details.
- Reasonable differences in pose, action, viewpoint, scale, composition, occlusion, and scene lighting must not be written as failure conditions.
- Do not assess whether the edit occurs on the correct person, at the correct position, in the correct quantity, or during the correct time span; that belongs to Instruction Following.
- Do not assess whether unedited regions remain unchanged; that belongs to Video Consistency.
- Do not assess generic video quality or physical naturalness.
- Do not construct questions that lack observable criteria, such as whether the result is exactly identical or perfectly replicated.

Output Requirements

All questions in `CheckList` must be in English.
Assign each question an integer `weight` indicating its relative importance within the current edit: `3` denotes a core constraint, `2` an important constraint, and `1` an auxiliary constraint.
Output only one valid JSON object containing only the list field `CheckList`. Do not output Markdown, explanations, answers, scores, or any additional text.

```
{
  "CheckList":  [
    {
      "question":  "A specific question answerable with yes or no
from the reference image and edited video",
      "weight":  3
    }
  ]
```

}

This is the source video before editing.
<SOURCE_VIDEO>

These are all editing instructions for this case, numbered in order.
Editing Instruction 1: <EDITING_INSTRUCTION_1>
Editing Instruction 2: <EDITING_INSTRUCTION_2>
⋮
Editing Instruction N: <EDITING_INSTRUCTION_N>

This is Reference Image 1.
<REFERENCE_IMAGE_1>

This is Reference Image 2.
<REFERENCE_IMAGE_2>
⋮
This is Reference Image N.
<REFERENCE_IMAGE_N>

Prompt for Video Consistency Checklist Generation

You are a strict, evidence-driven checklist designer for video editing fidelity.
Task
For the entire case, construct no more than five binary questions to evaluate whether the edited video changes only the content permitted by all editing instructions and preserves important information in the source video that should remain unchanged.
You will receive:

1. the source video;
2. all editing instructions for this case, numbered as "Editing Instruction N"; and
3. all reference images, where Reference Image N corresponds one-to-one with Editing Instruction N.

Analysis Procedure

1. Watch the complete source video to identify the main people, objects, backgrounds, text, actions, interactions, camera motion, editing structure, and event progression.
2. Consider all editing instructions for the case together and establish an "allowed changes list," including:
   - subjects, objects, clothing, attributes, backgrounds, styles, or VFX explicitly specified for addition, replacement, or modification; and
   - local occlusions, contact, shadows, reflections, or background completion necessary to perform those edits.
3. Establish a "must-preserve list," including:
   - people, objects, text, logos, and scene content not specified by any instruction;
   - aspects of an edited target's identity, appearance, and structure that are not specified for modification;
   - actions, poses, expressions, positions, quantities, scales, orientations, trajectories, interactions, and event timing not requested for modification; and
   - viewpoints, composition, camera motion, editing structure, and duration not requested for modification.
4. From the content that must be preserved, select no more than five of the most important, most susceptible to editing, and directly observable checks. Prioritize:
   - preservation of major non-target entities;
   - preservation of unspecified attributes of the edited target;
   - preservation of core actions, interactions, trajectories, and event order;
   - preservation of backgrounds, text, logos, and scene structure; and
   - preservation of shot, editing, and temporal structure.
5. Select questions adaptively for the current case. Do not add categories that need not be covered merely to reach five questions.

Boundaries for Different Editing Types

- **Subject or object replacement:** Changes to the target identity and specified appearance are allowed; the original target's action, position, scale, orientation, interaction, and time span of appearance should be preserved unless otherwise specified.
- **Subject or object addition:** The new content and its necessary occlusions and interactions are allowed; existing content must not be unjustifiably removed or replaced as a result.
- **Clothing replacement:** Changes to the target clothing are allowed; the person's identity, face, hairstyle, action, and other unspecified content should be preserved.
- **Local attribute modification:** Only the specified region or attribute may change; the person's identity, other clothing, object structure, or scene must not change as a side effect.
- **Background replacement:** Changes to the background are allowed; the foreground subjects, actions, relative layout, and shot structure should still be preserved.
- **Style transfer:** Changes to global color, lines, texture, material appearance, and rendering method are allowed; subject identity, scene semantics, actions, composition, and event progression should still be preserved.
- **VFX addition:** The effect and its necessary local changes in lighting, reflections, and occlusion are allowed; the host object and surrounding scene must not undergo unrelated structural changes.

Question Requirements

- The current case belongs to scenario category `<TYPE>`: `<TYPE_DEFINITION_AND_FOCUS>`. When constructing questions, focus on protecting content that should remain unchanged for this category, but do not treat changes permitted or required by any instruction as consistency failures.
- Each question must be answerable with `yes` or `no` by comparing the source video with the edited video provided later.
- Phrase every question positively: `yes` must always indicate that the content that should be preserved is preserved.
- Each question must refer to a specific person, object, action, scene, or shot visible in the current case. Do not merely ask whether other content remains unchanged.
- Each question must explicitly exclude all changes permitted by the instructions for the case. Do not treat an intended edit as a consistency failure.
- In principle, each question should check one preservation category. Avoid combining many different facts about people, backgrounds, actions, and shots into one question.
- Do not assess whether the target edit succeeds. Even if the model does not execute the instruction, the consistency item may still be satisfied as long as the original content is not rewritten in unrelated ways.
- Do not assess whether the edited content is similar to the reference image.
- Do not assess generic technical quality such as blur, compression, noise, or flicker.
- Do not assess whether interpenetration, floating, anatomy, or physical causality appears natural.
- Do not require pixel-level consistency. Allow encoding differences, natural motion blur, and minor perturbations that do not alter semantics or structure.

Output Requirements

All questions in `CheckList` must be in English.
Assign each question an integer `weight` indicating its relative importance within the current case: `3` denotes a core constraint, `2` an important constraint, and `1` an auxiliary constraint.
Output only one valid JSON object containing only the list field `CheckList`. Do not output Markdown, explanations, answers, scores, or any additional text.

```
{
  "CheckList":  [
    {
      "question":  "A specific question answerable with yes or no by
comparing the source and edited videos",
      "weight":  3
    }
  ]
}
```

This is the source video before editing.

<SOURCE_VIDEO>
These are all editing instructions for this case, numbered in order.
Editing Instruction 1: <EDITING_INSTRUCTION_1>
Editing Instruction 2: <EDITING_INSTRUCTION_2>
⋮
Editing Instruction N: <EDITING_INSTRUCTION_N>
This is Reference Image 1.
<REFERENCE_IMAGE_1>
This is Reference Image 2.
<REFERENCE_IMAGE_2>
⋮
This is Reference Image N.
<REFERENCE_IMAGE_N>

At runtime, <SOURCE_VIDEO>, <EDITING_INSTRUCTION_i>, and <REFERENCE_IMAGE_i> are instantiated with the source video, the i-th editing instruction, and its corresponding reference image, respectively. <TYPE> and <TYPE_DEFINITION_AND_FOCUS> are replaced by the scenario category and its corresponding dimension-specific definition and evaluation focus listed in Table 9. For Instruction Following and Reference Consistency, <TARGET_EDIT_INDEX> and <TARGET_INSTRUCTION> are additionally replaced by the one-based index and content of the target editing instruction.

Table 9: Scenario-category-specific definitions and evaluation focuses used to instantiate the checklist-generation prompts.

Scenario Category	Evaluation Dimension	Definition and Evaluation Focus
Basic Editing	Instruction Following	Assesses basic addition, replacement, or attribute modification, focusing on the editing operation, target, result, and any explicitly specified position, quantity, and scope.
	Reference Consistency	Focuses on the key visual features of the subject, object, clothing, attribute, background, style, or VFX referenced by the instruction.
	Video Consistency	Focuses on whether non-target content and the unspecified identity, structure, actions, and attributes of the edited target are preserved.
Target Disambiguation	Instruction Following	Requires localizing the edit target among similar entities based on visual features, action events, character identity, or spatial orientation, focusing on correct target localization and editing operation.
	Reference Consistency	Focuses on the reference visual features invoked by the instruction without reassessing whether target localization is correct.
	Video Consistency	Focuses on whether similar non-target entities are preserved and whether the edit incorrectly propagates to other entities.
Fast Movement	Instruction Following	The edit target undergoes fast or complex motion. Focuses on whether the edit remains associated with the correct target through motion, occlusion, and reappearance.
	Reference Consistency	Focuses on whether the reference identity and visual features remain stable through fast motion, pose and viewpoint changes, and reappearance.
	Video Consistency	Focuses on whether the original actions, poses, trajectories, speed, interactions, event timing, and camera tracking are preserved at the corresponding times.

Table 9 continued.

Scenario Category	Evaluation Dimension	Definition and Evaluation Focus
Inconspicuous Entities	Instruction Following	The edit target is small, distant, low-contrast, peripheral, occluded, or briefly visible. Focuses on whether the correct inconspicuous target is edited.
	Reference Consistency	Focuses on the category, silhouette, dominant color, and key structure discernible at normal video playback scale.
	Video Consistency	Focuses on whether the background texture around the target, nearby objects, and other small entities are preserved.
Multi-Shot Editing	Instruction Following	The source video contains multiple shots. Focuses on whether the edit covers all required shots or is restricted to the specified shots.
	Reference Consistency	Focuses on whether the reference identity and visual features remain consistent across relevant shots, shot scales, and viewpoints.
	Video Consistency	Focuses on whether the original number and order of shots, transitions, composition, event continuity, and non-target shots are preserved.
World Knowledge	Instruction Following	Requires physical laws, identity knowledge, semantic relationships, or contextual conditions to complete the edit, focusing on whether the necessary reasoning outcome is realized.
	Reference Consistency	Focuses on the reference identity, category, structure, and material features and, under reasonable state changes, whether these features remain traceable.
	Video Consistency	Focuses on whether subjects, environments, events, and shots beyond the necessary physical, semantic, and contextual changes are preserved.
Text Fidelity	Instruction Following	Requires adding or replacing reference text, focusing on whether the text appears on the correct carrier, in the correct region and shots, and with the correct orientation, perspective, and attachment.
	Reference Consistency	Checks the characters, spelling, numbers, and symbols of the reference text character by character, together with its reading order, arrangement, and visible visual style.
	Video Consistency	Focuses on whether original text, numbers, logos, signs, and layout positions not modified by the instruction are preserved.

H.2 Evaluation Prompts

We present the checklist-based evaluation prompts for Instruction Following, Reference Consistency, and Video Consistency, along with the evaluation prompts and scoring rubrics for Video Quality and Video Realism.

Prompt for Checklist-Based Evaluation

You are a strict, evidence-driven evaluator. Answer the supplied fixed checklist and output only the requested JSON object with lowercase `yes` or `no` answers.
This is the source video before editing.
`<SOURCE_VIDEO>`
This is the edited video.
`<EDITED_VIDEO>`
These are all editing instructions for this case, numbered in order.
Editing Instruction 1: `<EDITING_INSTRUCTION_1>`
Editing Instruction 2: `<EDITING_INSTRUCTION_2>`
⋮

Editing Instruction N: <EDITING_INSTRUCTION_N>

This is Reference Image 1.
<REFERENCE_IMAGE_1>

This is Reference Image 2.
<REFERENCE_IMAGE_2>

⋮

This is Reference Image N.
<REFERENCE_IMAGE_N>

Answer the fixed checklist for <EVALUATION_SCOPE>.
For each question, answer `yes` only when the supplied multimodal evidence clearly satisfies it. Otherwise answer `no`. Evaluate every question independently, preserve the given order, and do not add or remove questions.

Checklist

1. <CHECKLIST_QUESTION_1>
2. <CHECKLIST_QUESTION_2>

⋮

N. <CHECKLIST_QUESTION_N>

Output

Return exactly one JSON object containing only `checklist_results`. Its value must be a list with exactly <CHECKLIST_SIZE> lowercase `yes` or `no` strings in checklist order. Do not output reasons, scores, Markdown, or any additional text.

Prompt for Video Quality Evaluation

You are a strict, evidence-driven evaluator of technical video quality. Inspect each criterion internally, but output only one valid JSON object without the reasoning process, Markdown, or any additional text.
This is the video to be evaluated.
<VIDEO>

Task

Based only on the video to be evaluated, assess its technical and perceptual visual quality: whether the visual information is clear, complete, and stable, and whether it is degraded by encoding or generation artifacts.

Internal Inspection Procedure

Watch the complete video and inspect:

1. **Spatial clarity and detail:** unintended defocus or blur, low resolution, smeared details, oversharpening, aliasing, moiré patterns, banding, noise, pixelation, and block compression.
2. **Local integrity:** broken edges, discontinuous textures, stretching, ghosting, trails, anomalous holes, local melting, and foreground-background seams.
3. **Temporal stability:** flicker, texture crawling, details appearing and disappearing, local jumps, frame-to-frame jitter, motion trails, and persistent deformation in moving regions.
4. **Lighting, color, and visibility:** severe overexposure or underexposure, black or highlight clipping, abnormal color casts, uncontrolled contrast, and the resulting information loss.
5. **Playback integrity:** black frames, corrupted frames, unintended freezing, sudden collapse, or obvious discontinuities.

First identify the most salient defect, then assess its severity, spatial extent, and duration. Persistent or recurring problems affecting the main subject weigh more heavily than single-frame, peripheral, or small-region problems; do not simply add up the number of defects.

Dimension Boundaries

- Assess only technical visual quality, not narrative, content richness, compositional aesthetics, or creativity.
- Do not penalize cartoons, 2D or 3D animation, pixel art, hand-drawn imagery, film grain, or other styles themselves; penalize only unintended damage or instability within the given style.

- Reasonable depth of field, natural motion blur, camera motion, editing, grain, and lighting effects are not defects.
- Whether objects obey physical laws, actions are credible, or anatomy is reasonable belongs to Video Realism. Do not penalize these again unless they manifest as clear visual breakage, melting, flicker, or compositing artifacts.
- Use only visible evidence in the video; do not infer resolution settings, bitrate, the generation model, or the production source.

Scoring Rubric

- **5:** The video is clear, complete, and temporally stable, with almost no visible technical defects; fine-detail and moving regions are both well preserved.
- **4:** Overall quality is good, with only a few minor, local, or brief defects that do not impede viewing or content recognition.
- **3:** Several clearly visible moderate defects or one noticeable persistent problem are present; the main subject remains recognizable and viewing is largely unobstructed.
- **2:** Severe or persistent defects clearly impair subject detail, motion viewing, or visual integrity.
- **1:** Large-scale or prolonged collapse makes key information difficult to recognize, or the video contains severe corrupted frames, freezing, or visual destruction.

Output

The `reason` must be in Chinese and identify the primary defect, its spatial extent or duration, and its effect on viewing; for a high score, state which key quality aspects remain stable. Use one to three sentences. Output only:

```
{
  "reason":  "A concise justification based on observable technical
quality across the complete video",
  "quality_score":  5
}
```

Prompt for Video Realism Evaluation

You are a strict, evidence-driven evaluator of perceptual video realism. Inspect each criterion internally, but output only one valid JSON object without the reasoning process, Markdown, or any additional text. This is the video to be evaluated.
<VIDEO>

Task

Based only on the video to be evaluated, determine whether its visual world is natural, coherent, and credible within its own style. Here, "realism" means perceptual and physical plausibility; it does not require photorealism and does not involve guessing whether the video was generated by AI.

Internal Inspection Procedure

Watch the complete video and inspect:

1. **Morphology and function:** whether human or animal anatomy, object parts, and rigid or deformable structures are reasonable, and whether illogical fusion, proliferation, disappearance, or functional misplacement occurs.
2. **Motion and dynamics:** whether trajectories, speed, acceleration, inertia, gravity, collisions, contact, grasping, support, and deformation are natural, and whether motion starts and stops with reasonable transitions.
3. **Spatial relationships:** whether scale, depth, perspective, occlusion, front-back ordering, attachment, and contact between subjects and the scene are credible, and whether interpenetration, floating, sliding, or sticking occurs.
4. **Lighting and materials:** whether light sources, shadows, reflections, transparency, surface materials, and local color change plausibly with position, viewpoint, and motion.
5. **Temporal and causal coherence:** whether identity, shape, and state remain continuous; whether event order, state transitions, and causal processes are valid; and whether outcomes precede their causes or unexplained jumps occur.
6. **Scene integration:** whether newly added or composited elements share the same perspective, lighting and color, grain, edges, motion, and interactions as the background.

First identify the clearest implausible evidence, then determine the overall rating based on its certainty, importance, and duration. Repeated or persistent problems affecting the core action or main subject weigh more heavily than single-frame, local, or uncertain problems.

Dimension Boundaries

- Valid productions may be live action, visual effects, stop motion, 2D or 3D animation, cartoons, or other stylized content. Judge them according to their own style and motion rules; do not penalize them for lacking photorealism.
- Fantastical subjects, supernatural events, exaggerated expressions, flying, knockback, limited animation, held frames, or other stylized motion do not automatically reduce realism. Assess only whether their structure is stable and their motion, interactions, and causality are coherent within the established world.
- Do not assess compositional aesthetics or creativity.
- Low resolution, compression, ordinary blur, noise, or minor flicker belongs to Video Quality. Do not penalize these again unless they directly make structure, motion, spatial relationships, or causality implausible.
- Do not misinterpret normal motion, editing, shot changes, motion blur, brief occlusion, leaving the frame, or sampling intervals as anomalies. A single blurry or ambiguous frame is insufficient to establish structural deformation, object disappearance, or a state jump; do not penalize when a reasonable explanation cannot be ruled out.

Scoring Rubric

- **5:** Highly credible within its style; morphology, motion, space, lighting, materials, and causality are coherent, with almost no clear evidence of unnaturalness.
- **4:** Overall natural and credible; only minor, local, or brief unnatural phenomena are present and do not disrupt the main event.
- **3:** The basic world and actions are acceptable, but several clear moderate anomalies or one persistent problem that does not completely destroy the content is present.
- **2:** The main subject, key interaction, or object state repeatedly exhibits clear implausible problems across multiple moments, substantially disrupting immersion.
- **1:** The main subject or core event undergoes large-scale, persistent, and unambiguous structural, physical, or spatiotemporal collapse over an extended period; a single ambiguous frame or an anomaly whose persistence cannot be confirmed is insufficient for a score of 1.

Output

The `reason` must be in Chinese and identify the most important credible or implausible evidence and its duration; for scores from 1 to 3, provide the approximate time span of the anomaly. Use one to three sentences. Output only:

```
{
  "reason":  "A concise justification based on observable perceptual
realism across the complete video",
  "realism_score":  5
}
```

At runtime, for checklist-based evaluation, `<SOURCE_VIDEO>`, `<EDITED_VIDEO>`, `<EDITING_INSTRUCTION_i>`, and `<REFERENCE_IMAGE_i>` are instantiated with the source video, edited video, i-th editing instruction, and its corresponding reference image, respectively. `<EVALUATION_SCOPE>` is set according to the evaluation dimension, while `<CHECKLIST_QUESTION_i>` and `<CHECKLIST_SIZE>` are replaced by the i-th fixed checklist question and the total number of questions. For Video Quality and Video Realism, `<VIDEO>` is instantiated with the video to be evaluated.