OMNIEDIT-BENCH: A COMPREHENSIVE BENCHMARK FOR INSTRUCTION-BASED VIDEO EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Instruction-based video editing (IVE) is an emerging field with rich application scenarios, yet evaluating the editing models remains a significant challenge. Existing evaluation benchmarks for IVE task suffer from two fundamental limitations. First, task coverage is narrow and largely inherited from image editing, focusing on frame-level spatial manipulations while neglecting video-specific dimensions. Second, current evaluation metrics fail to capture instruction fidelity, allowing models to achieve high scores despite incorrect edits due to the strong visual prior of the original video. To this end, we introduce a comprehensive and clearly structured benchmark for IVE task. Our benchmark systematically decomposes editing tasks along multiple video-specific dimensions, including spatial, temporal, audio, and reference-based editing, going beyond conventional frame-level formulations. Moreover, it explicitly distinguishes between explicit and implicit instructions, incorporating reasoning-based scenarios to better reflect real-world editing requirements. This design enables a more complete and fine-grained evaluation of model capabilities across diverse dimensions. Furthermore, we propose a systematic evaluation framework that decomposes editing quality into four complementary dimensions: accuracy, preservation, realism, and consistency with both human judges and state-of-the-art vision language model (VLM). To emphasize the central role of instruction fidelity, we further introduce an accuracy-aware penalty mechanism, where the scores of other dimensions are conditioned on accuracy. This design prevents misleading high scores for visually plausible yet incorrect edits. We conducted experiments evaluating current prominent instruction-based video editing models, comprising both open-source and commercial models. Experimental results reveal that current models remain far from satisfactory in instruction-based video editing. `OmniEdit-Bench` provides a comprehensive and reliable testbed for instruction-based video editing, offering valuable insights into current model limitations and guiding future research in this direction.

1 INTRODUCTION

Instruction-based video editing (IVE) has emerged as a promising paradigm for controllable video manipulation. With the rapid advances in generative modeling (Ho et al., 2020; Song et al., 2020; Rombach et al., 2022; Song et al., 2021) and multimodal understanding (Yang et al., 2024), instruction-based video editing has seen remarkable progress (Zi et al., 2025; Bai et al., 2025; Cong et al., 2025; Wei et al., 2025), enabling diverse operations such as object removal (Miao et al., 2025; Zhou et al., 2023; Li et al., 2025; Jiang et al., 2025) and style transformation (Isola et al., 2017; Zhu et al., 2017; Zhang et al., 2023), and significantly advancing applications in video production.

Despite the rapid progress, evaluating IVE models remains a critical yet underexplored problem. The difficulty of evaluating instruction-based video editing stems from the inherently multi-dimensional nature of video content, which fundamentally distinguishes it from image editing (Chen et al., 2024; 2025a; Ruiz et al., 2023; Brooks et al., 2023; Hertz et al., 2022). Beyond spatial dimension, video editing involves additional dimensions such as temporal dynamics and audio signals. However, existing benchmarks (Chen et al., 2025b;c; Sun et al., 2025; Mou et al., 2025) fail to adequately account for this multi-dimensional complexity. Their task coverage remains narrow and largely inherited from image editing, focusing on frame-level spatial manipulations while neglecting video-specific dimensions such as temporal dynamics and audio (Tan et al., 2024; Oord et al., 2016; Kong

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 our evaluation tracks.** This figure illustrates representative tasks from **Spatial**, **Temporal**, **Audio**, **Reference**, and **Reasoning** tracks. By covering a wide range of scenarios, our benchmarks provide a comprehensive and precise basis for evaluating video editing models.

et al., 2020). Moreover, current evaluation metrics (Zhang et al., 2018; Wang et al., 2004; Hore & Ziou, 2010; Huang et al., 2024; Hessel et al., 2021) are substantially misaligned with the core objective of instruction-based editing. Metrics adapted by current editing models fail to reflect instruction fidelity, allowing models to achieve high scores despite incorrect edits due to the strong visual prior of the original video.

To this end, we introduce `OmniEdit-Bench`, a comprehensive, well-categorized benchmark tailored for evaluating instruction-based video editing capabilities. Our design explicitly accounts for the inherited multi-dimensional nature of video content, organizing tasks to reflect diverse editing requirements beyond static spatial changes. Specifically, we deliberately divides the benchmark into spatial, temporal, audio and reference-based track. Each track captures a unique facet of how videos can be manipulated under instructions. Furthermore, we structure our benchmark along the axis of instruction complexity. Specifically, we distinguish between explicit instructions, which directly specify the desired edits, and implicit instructions, which require models to infer user intent from

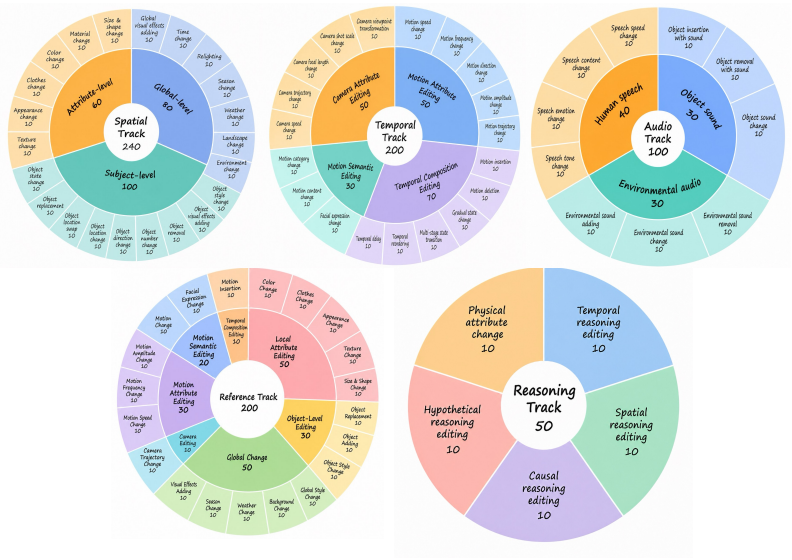

Figure 2: **Taxonomy of video editing tasks.** We classify the benchmark into five tracks: Spatial (240), Temporal (200), Reference (200), Audio (100), and Reasoning (50), illustrating the diverse scope of instruction-based video manipulation.

context. Within the implicit setting, we include reasoning-based scenarios that involve multi-step or compositional transformations across one or multiple dimensions. By organizing along these two complementary axes, `OmniEdit-Bench` enables a comprehensive and diagnostic evaluation.

For evaluation, we decompose the editing quality of videos into four dimensions: accuracy, preservation, realism, and consistency. This decomposition reflects the inherently multi-faceted nature of instruction-based video editing, where a successful edit must not only correctly execute the intended modification, but also preserve irrelevant content, maintain visual plausibility, and ensure coherence across frames and modalities. Importantly, these dimensions are not equally informative in isolation. In particular, high visual preservation quality or temporal consistency does not necessarily imply correct execution of the intended edit. To address this, we enforce instruction fidelity as a prerequisite for meaningful evaluation by introducing an accuracy-aware scoring mechanism, in which the scores of other dimensions are conditioned on accuracy. Specifically, the scores of the remaining dimensions are proportionally scaled by the accuracy score, reducing their contributions when accuracy is low. This design prevents visually plausible yet incorrect edits from receiving misleading high scores, leading to a more reliable and faithful assessment.

Using `OmniEdit-Bench`, we conduct a comprehensive evaluation of state-of-the-art instruction-based video editing models comprising both proprietary and open-source models. Our evaluation reveals consistent performance gaps across both open-source and commercial models. All models perform poorly on video-specific tracks such as temporal and audio as well as on more challenging reasoning-based scenarios, highlighting substantial room for improvement in handling multi-dimensional and implicit instructions. Horizontally, on relatively well-studied tracks such as spatial track, commercial models such as KlingV3-Omni (Team et al., 2025) and Seedance2.0 (Seedance et al., 2026)consistently achieve higher quantitative scores than open-source counterparts. This performance gap is also evident in qualitative aspects, including higher visual fidelity and longer temporal extent, reflecting advantages in data scale and system optimization.

2 OmniEdit-Bench

2.1 Benchmark Construction

To comprehensively evaluate IVE tasks, we construct `OmniEdit-Bench` as a multi-dimensional and systematically organized benchmark. As illustrated in Fig. 1 and Fig. 2, our design decomposes editing tasks along complementary axes, covering explicit and implicit instruction settings as well as diverse video-specific dimensions. Specifically, the benchmark includes five tracks: **spatial**, **tem-**

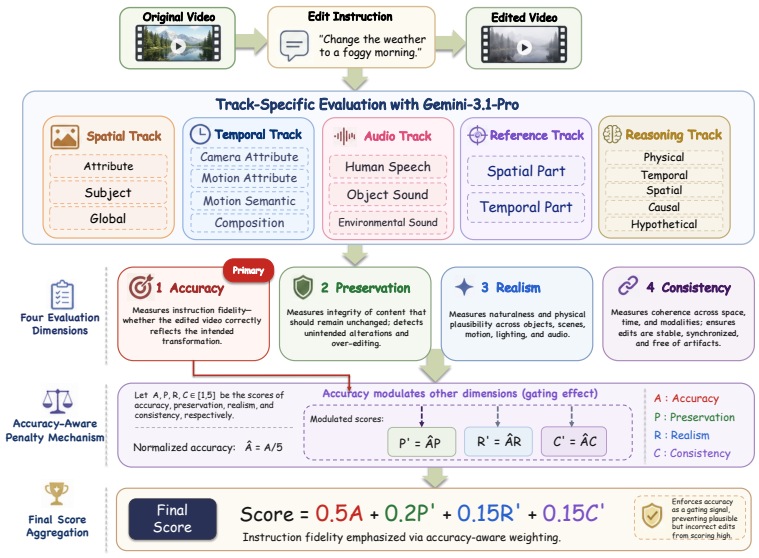

Figure 3: **Evaluation framework.** By introducing an accuracy-aware penalty mechanism, our framework ensures a precise assessment of video editing quality. The accuracy score acts as a gating signal, preventing plausible-looking but instruction-deviant edits from receiving high scores.

poral, **audio**, **reference-based**, and **reasoning**, each targeting a distinct aspect of editing capability. This design enables fine-grained and diagnostic evaluation of models across levels of complexity, ranging from low-level visual manipulation to high-level reasoning and multimodal alignment.

Spatial Track. The spatial track focuses on scenarios with limited motion, where consecutive frames remain highly similar. This setting largely aligns with image editing (Chen et al., 2024; 2025a) and is designed to evaluate models' ability to perform fine-grained visual modifications without substantial temporal variation. To systematically analyze different levels of editing granularity, we divide this track into three sub-tracks: attribute-level, subject-level, and global-level editing.

Attribute-level editing targets low-level fine-grained changes, such as color, texture, or material, and is designed to evaluate whether models can precisely modify local visual properties.

Subject-level editing focuses on object-centric changes, including removal, replacement, spatial manipulations or style changes . This category is defined by an intermediate level of granularity, where the scope of editing expands from local attribute changes to semantically entire objects.

Global-level editing involves scene-wide transformations, such as relighting, weather or season changes, and represents the coarsest level of granularity. This category evaluates whether models can perform coherent transformations that affect the scene as a whole rather than isolated regions.

Temporal Track. Temporal track is designed to capture the intrinsic temporal complexity of video editing, which is mostly absent in previous benchmarks. This track focuses on videos with significant camera or object motion and we organize this track into four sub-tracks: camera attribute editing, motion attribute editing, motion semantic editing, and temporal composition editing.

Camera attribute editing targets transformations in camera parameters, such as viewpoint, shot scale, and trajectory. These operations require modeling continuous camera motion over time, introducing long-range temporal dependencies across frames. Models must ensure smooth transitions, consistent motion trajectories, and stable geometric relationships throughout the video.

Motion attribute editing focuses on controlling quantitative and basic properties of motion, such as speed, direction, and amplitude. Unlike camera attribute editing modifying the observation perspective, this category directly alters object dynamics within the scene. This sub-track evaluates whether models can precisely modulate motion signals while maintaining coherent temporal evolution.

Motion semantic editing involves higher-level transformations of motion content or category, such as altering the type or meaning of an action. This category requires models to understand and recompose semantically meaningful motion patterns over time. Such edits span multiple frames and

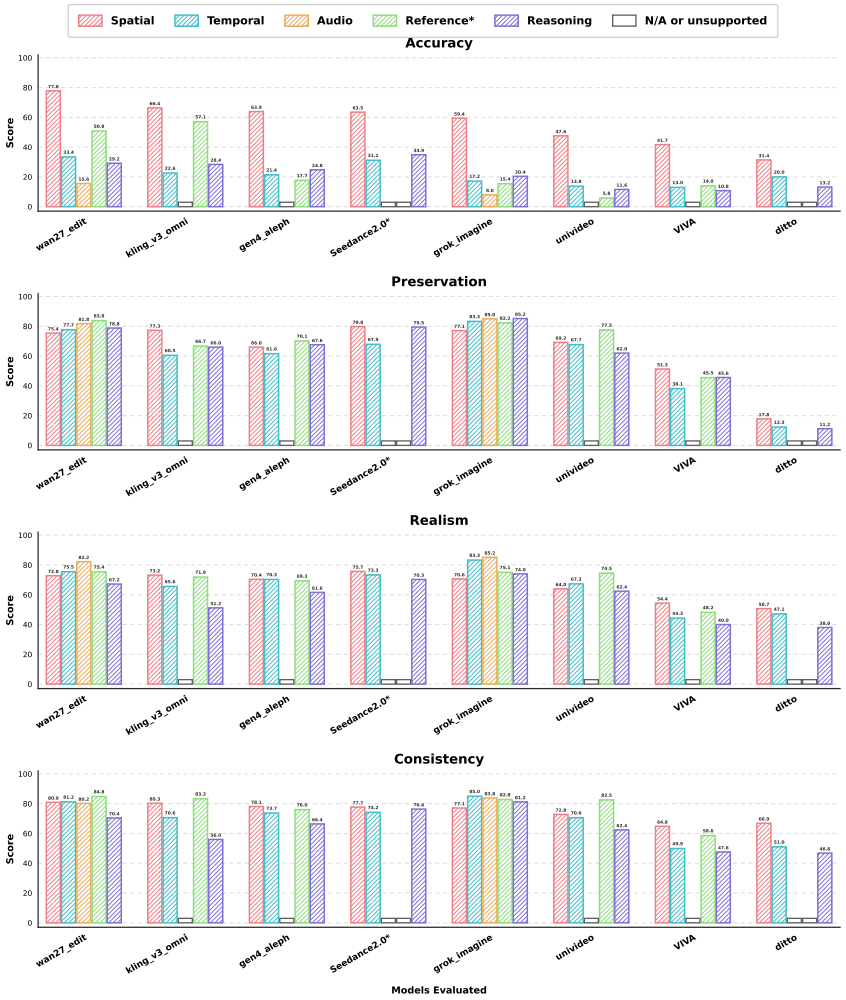

Figure 4: **Comparisons of model performance across evaluation metrics.** The bar charts display the score distribution of all the evaluated models for Accuracy, Preservation, Realism, and Consistency, categorized by the five tracks defined in OmniEdit-Bench. *Note: For Reference track, only models with spatial reference capabilities are evaluated due to current architectural limitations.*

involve structured temporal dependencies, where the global motion pattern must remain coherent while being semantically transformed. This sub-track evaluates whether models can perform temporally consistent motion reinterpretation rather than merely adjusting low-level motion statistics.

Temporal composition editing captures more complex operations over entire motion sequences, including motion insertion, reordering, and synchronization. These operations introduce long-range temporal dependencies and often involve multi-step transformations, where local consistency must be preserved while enforcing new temporal relationships across distant frames. This makes temporal composition editing particularly challenging, as models must coordinate motion events over the video and avoid inconsistencies such as temporal misalignment or incoherent transitions.

Audio Track. Video inherently involves visual and auditory signals, making audio a critical yet underexplored dimension for IVE task. Audio track evaluates whether models can perform coherent audio-visual editing, where modifications to the visual content must be accompanied by temporally aligned and semantically consistent changes in audio. This track is divided into three sub-tracks: human speech editing, object sound editing, and environmental audio editing.

Human speech editing focuses on modifications related to human speech, including changes in speech content, speaking speed, emotion, and tone. This category evaluates whether models can generate temporally aligned speech signals that are consistent with the visual context, such as lip movements and speaker identity, while preserving natural prosody and semantic correctness.

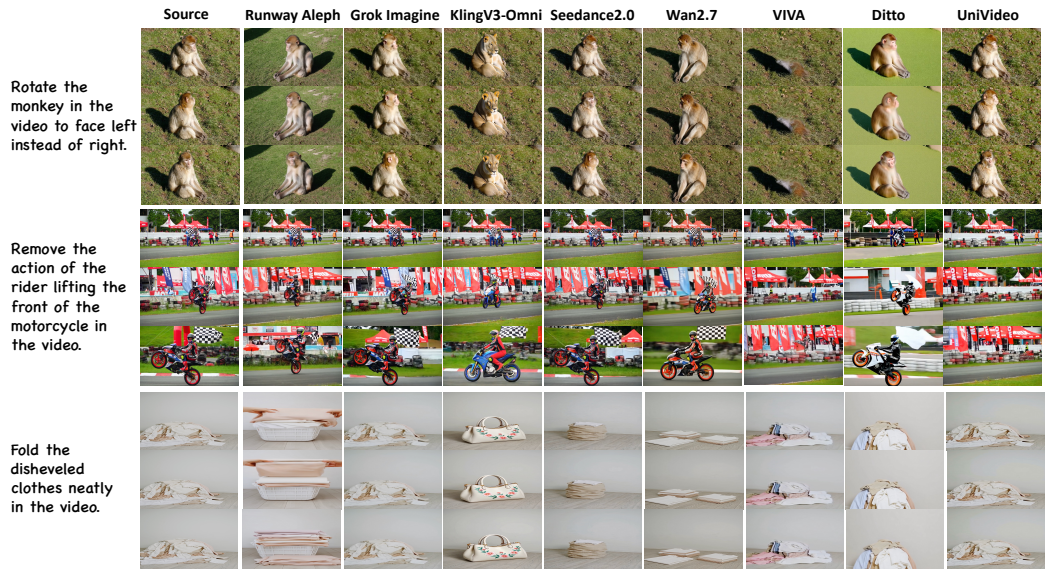

Figure 5: **Some examples of models' output on `OmniEdit-Bench`.** Zoom in for better view. For more visualization results, please refer to Appx. G.

Object sound editing addresses audio changes associated with object-level interactions. Typical tasks include inserting or removing objects along with their sounds, or modifying the sounds produced by objects. This category focuses more on general objects inluding animals and vehicles and requires models to maintain consistency between visual events and their associated sounds.

Environmental audio editing focuses on background or ambient sounds, such as adding, modifying, or removing environmental audio. These edits require maintaining global audio consistency over time while ensuring compatibility with the visual scene and the corresponding editing instruction.

Reference-based Track. We further introduce a reference-based track to evaluate models' ability to incorporate external visual guidance during editing. Rather than defining a completely independent set of tasks, this track is built upon both spatial and temporal editing scenarios by introducing reference inputs, requiring models to perform edits that are not only instruction-following but also aligned with a given reference. In the spatial setting, reference-based tasks are instantiated using both reference images and reference videos, where both forms primarily provide spatial guidance. The temporal setting relies exclusively on reference videos, as dynamic transformations such as motion patterns and camera trajectories can not be adequately specified by a single image. We further organize reference-based tasks into spatial and temporal editing, evaluating whether models can perform edits while faithfully aligning with reference signals across both dimensions.

Reasoning Track. Reasoning track is designed with using implicit editing instructions where the desired editing outcome is not directly specified but must be inferred from the instruction. These tasks require models to interpret user intent and perform multi-step reasoning before generating the edited video. We organize these tasks into five categories, each capturing a distinct type of reasoning.

Physical attribute change involves modifying underlying physical properties in the scene, such as gravity, friction, or weight, requiring models to reason about physically plausible outcomes.

Spatial reasoning editing focuses on inferring spatial arrangements and relationships, such as reorganizing a cluttered scene into a structured layout, demanding an understanding of object grouping, alignment, and spatial constraints.

Temporal reasoning editing requires models to infer time-dependent outcomes, such as predicting scene evolution after a given duration (e.g., traffic light changes after a certain time), emphasizing temporal progression and consistency.

Causal reasoning editing involves reasoning about cause-and-effect relationships, such as determining the state of a system given certain conditions (e.g., a solved Rubik's cube), requiring models to generate outcomes consistent with underlying causal logic.

Table 1: **Overall performance on OmniEdit-Bench.** We report the performance of various video editing models across five primary tracks. Higher scores indicate better alignment with the evaluation dimensions. *Note: Seedance 2.0 is only evaluated on Spatial, Temporal, and Reasoning tracks due to copyright restrictions. And for Reference track, only models with spatial reference capabilities are evaluated due to current architectural limitations.*

Models	Resolution	Spatial	Temporal	Audio	Reference*	Reasoning	Overall Score (Average)
Runway Aleph	720P	54.2	16.5	-	15.1	20.3	24.2
Grok Imagine	720P	51.7	15.0	6.7	13.1	17.9	19.0
KlingV3-Omni	720P	59.6	18.5	-	49.9	22.9	38.3
Seedance2.0*	720P	56.9	25.2	-	-	29.8	37.3
Wan2.7-Edit	720P	69.8	29.0	13.6	46.4	24.6	36.0
VIVA	480P	33.5	9.6	-	9.8	7.5	14.1
Ditto	480P	22.9	13.5	-	-	8.6	15.0
UniVideo	480P	39.9	11.3	-	4.9	8.4	14.4

Hypothetical reasoning editing addresses counterfactual or imaginative scenarios (e.g., "what if" conditions), where models must generate plausible alternative outcomes that are not directly observed, requiring abstraction beyond the given input.

2.2 Evaluation Pipeline

To systematically evaluate instruction-based video editing performance, as shown in Fig. 3, we design a multi-dimensional evaluation pipeline that decomposes editing quality into four complementary aspects: **accuracy**, **preservation**, **realism**, and **consistency**. Each dimension is independently assessed using a state-of-the-art vision-language model (VLM), specifically Gemini-3.1-Pro, with task-specific prompts tailored to different tracks (spatial, temporal, audio, reference, and reasoning). The detailed prompt designs for each track are illustrated in Appx. F.

Accuracy measures whether the edited video correctly reflects the intended transformation specified by the instruction. It serves as the primary indicator of instruction fidelity. The evaluation of accuracy varies across tracks to capture task-specific requirements. For spatial editing, accuracy is evaluated at different levels of granularity. For temporal editing, it further includes action content accuracy, motion attribute accuracy, timing accuracy, and camera control accuracy, reflecting both what is edited and when it occurs. For audio editing, accuracy considers both sound correctness and temporal alignment. In the reasoning track, accuracy evaluates whether the model successfully infers and executes implicit instructions, including spatial, temporal, and causal reasoning.

Preservation evaluates whether the edited video maintains the integrity of content that should remain unchanged. In IVE task, only part of the scene is expected to be modified, and unintended alterations to irrelevant regions indicate poor editing control. This dimension is particularly important for distinguishing precise editing from over-editing.

Realism measures whether the edited video appears natural and physically plausible. We evaluate realism at multiple levels. Physical realism assesses whether the edited content obeys physical constraints, such as lighting consistency and shadows. Subject realism examines whether the appearance of objects or characters remains plausible after editing. Environment realism evaluates the coherence of the overall scene. For temporal and motion-related tasks, we additionally consider motion realism, ensuring that dynamic behaviors appear natural rather than artificial. For audio-related tasks, realism also includes the naturalness of generated or modified audio signals.

Consistency evaluates whether the editing results are coherent. We consider three forms of consistency. Temporal consistency ensures that edits are smoothly and coherently applied across frames without flickering or discontinuities. Spatial consistency evaluates whether spatial relationships and structures remain stable after editing. For audio-related tasks, we additionally consider audio-visual consistency, ensuring that audio signals are synchronized with visual content.

Accuracy-Aware Penalty Mechanism. To emphasize instruction fidelity, as illustrated in Fig. 3, we introduce an accuracy-aware penalty mechanism that adjusts the contributions of different evaluation dimensions. Let $A, P, R, C \in [1, 5]$ denote the scores of accuracy, preservation, realism, and consistency, respectively, and define the normalized accuracy as $\hat{A} = A/5$. The scores of the remaining dimensions are proportionally modulated as $P' = \hat{A}P$, $R' = \hat{A}R$, and $C' = \hat{A}C$. The final score is then computed as a weighted combination: Score $= 0.5A + 0.2P' + 0.15R' + 0.15C'$.

Table 2: Ablation of the accuracy-aware scoring mechanism. EA: Equal Average; WA: Weighted Average; AAP: Accuracy-aware Penalty.

Method	Human Corr.↑	Pearson↑	MAE↓
EA	0.846	0.861	0.468
WA	0.889	0.894	0.392
AAP	**0.934**	**0.921**	**0.287**

Table 3: Sensitivity analysis of different weight configurations.

w_A	w_P	w_R	w_C	Human $\rho\uparrow$	MAE↓	Ranking $\tau\uparrow$
0.25	0.25	0.25	0.25	0.881	0.374	0.91
0.35	0.25	0.20	0.20	0.907	0.341	0.94
0.50	0.25	0.125	0.125	0.928	0.301	**0.98**
0.50	**0.20**	**0.15**	**0.15**	**0.934**	**0.287**	**0.98**
0.50	0.15	0.25	0.10	0.919	0.315	0.96
0.60	0.16	0.12	0.12	0.925	0.306	**0.98**
0.70	0.12	0.09	0.09	0.910	0.329	0.94

In this formulation, accuracy serves as a controlling factor that continuously scales the influence of other dimensions, ensuring that their contributions diminish when the editing accuracy is low. This design prevents visually plausible but incorrect edits from receiving inflated scores, while still allowing high-quality edits to benefit from strong performance across all dimensions.

3 Experiments

3.1 Main Results

We evaluate a diverse set of state-of-the-art instruction-based video editing models on `OmniEdit-Bench`, covering both open-source and commercial systems. The results are summarized in Tab. 1, where we report performance on a 100-point scale across five tracks. To further analyze model behavior beyond the aggregated scores, we report dimension-wise evaluation results in Fig. 4 across different tracks and models. In addition, to validate the reliability of our evaluation protocol, we compare VLM-based scores with human annotations, as shown in Tab. 4.

3.2 Comparison across Models

Tab. 1 reveals clear performance differences across models. Overall, commercial systems consistently outperform open-source models, particularly on well-studied tasks such as spatial editing. However, this advantage diminishes on more challenging tracks. Even the strongest models show limited performance on temporal and reasoning tasks. For example, while Wan2.7-Edit reaches 29.0 on the temporal track, most other models remain below 20, and reasoning performance remains modest overall, with the best score (Seedance2.0) only reaching 29.8. This suggests that improvements in model scale and training data primarily benefit spatial tasks, but are insufficient for addressing more complex temporal and reasoning challenges. On the reference track, performance varies significantly across models, with KlingV3-Omni (49.9) and Wan2.7-Edit (46.4) outperforming others by a large margin, reflecting stronger capabilities in reference-conditioned generation. Open-source models, while generally lagging behind, still exhibit competitive performance in certain settings. For example, UniVideo achieves a reasonable spatial score (39.9), indicating potential for further improvement with better training strategies and architectural design.

3.3 Comparisons across Tracks

From a task perspective, we observe consistent performance disparities across different tracks. While models achieve relatively strong results on the spatial track (e.g., up to 69.8), their performance drops significantly on temporal, audio, and reasoning tracks. The temporal track remains particularly challenging, with most models scoring below 20 and only a few exceeding 25. This reflects the difficulty of maintaining coherent motion and modeling long-range temporal dependencies across frames. The reasoning track further exposes limitations in instruction understanding. Despite being a core component of realistic editing scenarios, reasoning scores remain relatively low across all models (mostly below 25), indicating that current systems struggle to infer implicit instructions. The audio track highlights a critical gap in multimodal capabilities. Only a subset of models supports audio editing, and overall performance remains low, confirming that audio remains an underexplored dimension in video editing. These results suggest that while current models perform relatively well on spatial editing, they remain far from handling realistic video editing scenarios.

Table 4: **Alignment between VLM-based evaluation and human judgments.** For each VLM score level (1–5), we report the proportion of samples, the mean and standard deviation of human scores, and the mean absolute error (MAE), across four evaluation dimensions: Accuracy (Acc.), Preservation (Pres.), Realism (Real.), and Consistency (Cons.).

Score	Prop. (%)				Human Mean				Human Std.				MAE			
	Acc.	Pres.	Real.	Cons.	Acc.	Pres.	Real.	Cons.	Acc.	Pres.	Real.	Cons.	Acc.	Pres.	Real.	Cons.
1	39.35	19.07	12.84	11.47	1.48	1.80	1.28	1.89	0.55	0.42	1.20	1.40	0.85	0.83	0.80	1.61
2	15.64	9.88	17.91	12.77	2.18	2.64	2.37	2.15	1.50	1.22	1.10	1.24	1.64	0.96	0.84	0.75
3	5.52	9.86	15.34	10.18	3.87	3.33	3.73	3.50	0.83	0.96	0.80	0.78	0.77	0.70	0.60	0.78
4	14.02	19.10	23.12	24.68	4.26	3.81	3.67	4.46	0.94	1.46	0.64	0.83	0.70	1.12	0.76	0.79
5	25.47	42.09	30.78	40.90	4.61	4.44	4.57	4.87	1.12	0.88	0.51	0.34	0.59	0.56	0.13	0.23
Overall	-	-	-	-	-	-	-	-	-	-	-	-	**0.86**	**0.77**	**0.55**	**0.64**

3.4 Human Alignment

To evaluate the reliability of our evaluation pipeline, we compare VLM-based scores with human judgments in Tab. 4. For each dimension, we group samples by the discrete scores (1–5) assigned by Gemini 3.1-pro and report the corresponding human score statistics, including mean, standard deviation, and mean absolute error (MAE). Human scores closely track VLM scores with a clear monotonic trend: samples receiving the highest VLM score (5) achieve high human ratings (4.61–4.87), while lower VLM scores correspond to lower human ratings. The overall MAEs are 0.86, 0.77, 0.55, and 0.64 for Accuracy, Preservation, Realism, and Consistency, respectively, indicating good VLM–human agreement. Human score variance also decreases at higher score levels, suggesting stronger agreement on high-quality edits. Overall, these results demonstrate that VLM-based evaluation provides a reliable and scalable proxy for human judgment. Further details on the human evaluation protocol, including annotator recruitment, calibration, and inter-annotator agreement, are provided in Appx. C.

3.5 Ablation and Sensitivity Analysis of Accuracy-aware Scoring

Effectiveness of Accuracy-aware Scoring. We first ablate the proposed accuracy-aware mechanism by comparing it with Equal Average and Weighted Average. As shown in Tab. 2, introducing dimension-specific weights improves Human Corr. from 0.846 to 0.889 and reduces MAE from 0.468 to 0.392. More importantly, incorporating the accuracy-aware penalty further improves Human Corr. to 0.934 and Pearson correlation to 0.921, while reducing MAE to 0.287. This demonstrates that the improvement does not merely come from assigning a larger weight to Accuracy; explicitly penalizing inaccurate edits better captures its prerequisite role in successful editing.

Sensitivity to Dimension Weights. We further analyze the sensitivity to different weight configurations in Tab. 3. Increasing the emphasis on Accuracy initially improves human alignment, with our adopted configuration $(0.50, 0.20, 0.15, 0.15)$ achieving the highest Human Spearman correlation (0.934), the lowest MAE (0.287), and a high ranking consistency ($\tau = 0.98$). However, further increasing the Accuracy weight degrades performance, indicating that Accuracy should be dominant rather than exclusive. These results support our design of treating Accuracy as the primary criterion while retaining Preservation, Realism, and Consistency as complementary dimensions.

4 Conclusion

In this paper, we introduce `OmniEdit-Bench`, a comprehensive and systematically structured benchmark for instruction-based video editing, covering five tracks: spatial, temporal, audio, reference-conditioned, and reasoning. We further propose a unified evaluation pipeline that decomposes editing quality into four dimensions: accuracy, preservation, realism, and consistency, together with an accuracy aware penalty mechanism to ensure faithful assessment of instruction fidelity. Extensive experiments show that while existing models perform well on spatial editing, their performance degrades substantially on more challenging tracks involving temporal coherence, multimodal alignment, and implicit reasoning, with even state-of-the-art commercial models exhibiting clear limitations. These findings highlight fundamental challenges in instruction-based video editing and point to important future directions.

AI USE STATEMENT

In this work, we used generative AI tools to synthesize a small portion of the benchmark data and employed vision-language models (VLMs) as part of our evaluation pipeline. We additionally used generative AI tools to assist with paper writing, including language refinement, proofreading, and readability improvement. All AI-generated or AI-assisted data, evaluation results, and manuscript content were reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, experimental results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

OmniEdit-Bench is constructed primarily from publicly accessible videos and public datasets, supplemented with a small number of synthetically generated videos for rare or difficult-to-collect scenarios. All source data are used under licenses that permit academic research, and we avoid collecting private or personally identifiable information. Human evaluation was conducted by trained annotators following a predefined evaluation protocol, and the collected ratings are reported only in aggregate without releasing annotator-specific information. As video editing models may potentially be misused to create misleading or manipulated content, we emphasize that this benchmark is intended solely for research purposes to facilitate the transparent and systematic evaluation of video editing systems.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the benchmark construction, task taxonomy, evaluation dimensions, and scoring protocol in the main paper. The appendix further documents the benchmark construction process, evaluation prompts, human evaluation protocol, ablation and sensitivity analyses, and additional statistical analyses. To facilitate reproducibility, we will release the benchmark data and annotations, evaluation code, and evaluation prompts upon publication, subject to the licenses and redistribution requirements of the original data sources. Together, these materials provide the necessary information to reproduce the benchmark evaluation and the main experimental results reported in this work.

REFERENCES

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. *arXiv preprint arXiv:2510.15742*, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *CVPR*, 2025a.

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. *arXiv preprint arXiv:2510.11647*, 2025b.

Yupeng Chen, Penglin Chen, Xiaoyu Zhang, Yixian Huang, and Qian Xie. Editboard: Towards a comprehensive evaluation benchmark for text-based video editing models. In *AAAI*, 2025c.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

Xiaoyan Cong, Haotian Yang, Angtian Wang, Yizhi Wang, Yiding Yang, Canyu Zhang, and Chongyang Ma. Viva: Vlm-guided instruction-based video editing with reward optimization. *arXiv preprint arXiv:2512.16906*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

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. *arXiv preprint arXiv:2512.07826*, 2025.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, 2010.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance. In *CVPR*, 2025.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *NeurIPS*, 2020.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. Diffueraser: A diffusion model for video inpainting. *arXiv preprint arXiv:2501.10018*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Xinyu Liu, Hangjie Yuan, Yujie Wei, Jiazheng Xing, Yujin Han, Jiahao Pan, Yanbiao Ma, Chi-Min Chan, Kang Zhao, Shiwei Zhang, et al. Revise: Towards reason-informed video editing in unified models with self-reflective learning. *arXiv preprint arXiv:2512.09924*, 2025.

Chenxuan Miao, Yutong Feng, Jianshu Zeng, Zixiang Gao, Liu Hantang, Yunfeng Yan, Donglian Qi, Xi Chen, Bin Wang, and Hengshuang Zhao. Rose: Remove objects with side effects in videos. In *NeurIPS*, 2025.

Chong Mou, Qichao Sun, Yanze Wu, Pengze Zhang, Xinghui Li, Fulong Ye, Songtao Zhao, and Qian He. Instructx: Towards unified visual editing with mllm guidance. *arXiv preprint arXiv:2510.08485*, 2025.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

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 *AAAI*, 2025.

Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *ECCV*, 2024.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

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

Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *TIP*, 2004.

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. *arXiv preprint arXiv:2510.08377*, 2025.

xAI. Grok imagine. `https://x.ai`, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *CVPR*, 2023.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025.

Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *ICCV*, 2023.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2m: A high-quality instruction-based dataset for general video editing by video specialists. In *NeurIPS*, 2025.

A Related Work

Instruction-based Video Editing. Instruction-based video editing (IVE) has recently emerged as a compelling paradigm. Early approaches (Guo et al., 2023) largely adapt image editing methods to the temporal domain by incorporating attention or consistency modules into text-to-image diffusion frameworks (Saharia et al., 2022; Peebles & Xie, 2023; Labs et al., 2025). While such extensions can preserve spatial structure, they remain limited in handling dynamic scenarios involving object motion or camera movement. More recent methods adopt architectures including video diffusion (Ho et al., 2022) and flow-matching models (Lipman et al., 2023; Liu et al., 2022), which improve temporal coherence and editing fidelity. Representative open-source systems include Señorita-2M (Zi et al., 2025), Ditto (Bai et al., 2025), VIVA (Cong et al., 2025), and UniVideo Wei et al. (2025), while commercial models such as Kling Omni (Team et al., 2025), Seedance2.0 (Seedance et al., 2026), Wan (Wan et al., 2025; Cheng et al., 2025), and Grok Imagine (xAI, 2024) have achieved strong empirical performance. Despite these advances, model capabilities remain uneven, with systems excelling on different editing types, underscoring the need for systematic and standardized evaluation.

Table 5: Comparison of OmniEdit-Bench with existing benchmarks across different editing dimensions. ✓ indicates support, ✗ indicates lack of coverage, and ✓✗ indicates partial support.

Benchmark	Spatial	Temporal	Audio	Reference	Reasoning
VE-Bench	✓	✗	✗	✗	✗
EditBoard	✓	✗	✗	✗	✗
OpenVE-3M	✓	✗	✗	✗	✗
IVE-Bench	✓	✓✗	✗	✗	✗
VIE-Bench	✓	✗	✗	✓	✗
RVE-Bench	✓	✗	✗	✗	✓
OmniEdit-Bench	✓	✓	✓	✓	✓

Benchmarks for Instruction-based Video Editing. In parallel with the development of IVE methods, several evaluation benchmarks have been proposed. Early efforts such as VE-Bench (Sun et al., 2025) and EditBoard (Chen et al., 2025c) focus on a limited set of tasks largely inherited from image editing, and remain constrained in both scale and diversity. Later benchmarks, including OpenVE-3M (He et al., 2025) and IVE-Bench (Chen et al., 2025b), expand the data coverage but do not explicitly capture video-specific aspects such as temporal and audio. Some works further explore individual dimensions, where VIE-Bench (Mou et al., 2025) studies reference-conditioned editing, and RISE-Bench (Zhao et al., 2025) together with RVE-Bench (Liu et al., 2025) investigate reasoning-based editing. However, these efforts remain fragmented and do not provide a unified evaluation framework. To address this gap, we propose `OmniEdit-Bench`, a comprehensive and structured benchmark for multi-dimensional evaluation of instruction-based video editing.

B Data Source and Categories of OmniEdit-Bench

The videos in `OmniEdit-Bench` are collected from three primary sources: (1) open-source video platforms, such as Pexels and Pixabay, which provide diverse real-world footage; (2) publicly available video datasets, including OpenVid-1M, offering large-scale and well-curated video content; and (3) state-of-the-art text-to-video generation models, which enable the inclusion of synthetic yet controllable scenarios that are difficult to obtain from real-world data.

To ensure both diversity and representativeness, we carefully curate videos spanning a wide range of categories, including plants, animals, human activities, natural environments, and human-made scenes. This diverse composition allows the benchmark to cover varied visual appearances, motion patterns, and interaction types. In addition, we explicitly balance the data across different content domains to avoid bias toward specific scene types.

C Human Evaluation Details

Annotators and Calibration. We recruited 30 trained undergraduate annotators majoring in Artificial Intelligence to conduct the human evaluation. The annotators were independent of the benchmark construction team to avoid potential bias introduced during data collection and instruction design. Before the formal evaluation, all annotators participated in a calibration session based on a unified annotation guideline. The guideline provides detailed definitions of the evaluation criteria, scoring rules, and representative examples for all five editing tracks (*Spatial*, *Temporal*, *Audio*, *Ref-*

Table 6: Inter-annotator agreement across the four evaluation dimensions.

Metric	Accuracy	Preservation	Realism	Consistency
Krippendorff's α	0.84	0.79	0.75	0.77
ICC(2,k)	0.91	0.87	0.84	0.86

Table 7: Human alignment and model-ranking consistency across different VLM evaluators.

Evaluator	Human $\rho \uparrow$	Human MAE $\downarrow$	Ranking $\tau \uparrow$
Gemini-2.5-Pro	**0.934**	**0.287**	**0.983**
Qwen2.5-VL-72B	0.918	0.315	0.978
InternVL3-78B	0.912	0.328	0.965

erence, and *Reasoning*). The calibration process was designed to ensure that annotators developed a consistent understanding of the evaluation dimensions and scoring criteria before assessing the edited videos.

Evaluation Protocol. Human evaluation follows the same four-dimensional evaluation protocol as our VLM-based evaluator, including *Accuracy*, *Preservation*, *Realism*, and *Consistency*. Each dimension is independently rated on a discrete scale from 0 to 5 according to the annotation guideline. Among these dimensions, *Accuracy* is treated as the primary criterion, measuring whether the requested edit is correctly executed, while *Preservation*, *Realism*, and *Consistency* serve as complementary criteria characterizing the overall quality of the edited video. Representative examples for different score levels and editing tracks were provided during calibration to reduce ambiguity and improve consistency across annotators.

To comprehensively validate the alignment between our automatic evaluator and human judgments, human annotators independently assessed the edited outputs generated by all evaluated models using the same evaluation protocol. In total, more than 4,500 edited samples were manually evaluated, covering all five editing tracks and all evaluated models. This large-scale human evaluation enables us to systematically analyze the agreement between human judgments and VLM-based evaluation at both the sample and model-ranking levels.

Inter-Annotator Agreement. We further measure inter-annotator agreement to assess the reliability of the collected human judgments. Specifically, we report both Krippendorff's α and the intraclass correlation coefficient (ICC) for each evaluation dimension. As shown in Tab. 6, the resulting agreement scores are consistently high across all four dimensions, indicating substantial agreement among independent annotators and supporting the reliability of our human evaluation results.

D Cross-VLM Evaluator Analysis

To examine whether our evaluation framework depends on a specific VLM, we further evaluate two representative open-source VLMs, Qwen2.5-VL-72B and InternVL3-78B, using the same evaluation prompts and scoring protocol as Gemini-2.5-Pro.

Cross-VLM Consistency and Human Alignment. As shown in Tab. 7, all three evaluators exhibit strong agreement with human judgments. Gemini-2.5-Pro achieves the highest Human Spearman correlation (0.934) and the lowest MAE (0.287), while Qwen2.5-VL-72B and InternVL3-78B also maintain high correlations of 0.918 and 0.912, respectively. Moreover, the resulting model rankings remain highly consistent with human rankings across all evaluators, with Kendall's $\tau \geq 0.965$. Together with pairwise Spearman correlations above 0.95 among the three evaluators, these results indicate that our evaluation framework is robust to the choice of VLM evaluator.

Human Alignment across Editing Tracks. We further examine human alignment separately for each editing track. As shown in Tab. 8, all three evaluators exhibit a consistent trend across tracks. Spatial and Reference editing show stronger agreement with human judgments, whereas Audio and particularly Reasoning remain more challenging. Nevertheless, the differences among evaluators

Table 8: Human alignment (MAE) of different VLM evaluators across editing tracks. Lower is better.

Track	Gemini ↓	Qwen ↓	InternVL ↓
Spatial	**0.241**	0.263	0.271
Temporal	**0.312**	0.334	0.348
Audio	**0.386**	0.417	0.432
Reference	**0.279**	0.301	0.318
Reasoning	**0.451**	0.486	0.503

Table 9: Bootstrap estimates of benchmark scores with 95% confidence intervals across different editing tracks.

Track	Mean	95% CI
Spatial	4.18	± 0.05
Temporal	4.06	± 0.06
Audio	3.82	± 0.10
Reference	4.01	± 0.06
Reasoning	3.57	± 0.14

remain small across all tracks, further demonstrating the robustness of our evaluation framework across diverse editing capabilities.

E Bootstrap Confidence Interval Analysis

To assess the statistical stability of our benchmark results, we perform bootstrap resampling to estimate 95% confidence intervals (CIs) for the benchmark scores. Specifically, benchmark instances are resampled with replacement within each editing track, and the corresponding scores are recomputed for each bootstrap sample.

As shown in Tab. 9, the confidence intervals remain narrow across all editing tracks. Spatial, Temporal, and Reference achieve particularly small intervals (± 0.05–± 0.06). Although Audio and Reasoning contain fewer benchmark instances, their intervals remain reasonably small at ± 0.10 and ± 0.14, respectively. These results indicate that the benchmark scores are statistically stable across different editing tracks.

F Prompts for VLM Evaluation

Base Instruction: You are a STRICT and PROFESSIONAL evaluator for video editing tasks. You will be given: Original Video, Edited Video, User Edit Instruction, and (Optional) Reference (image or description). Your task is to evaluate the Edited Video.Core principles:1. ACCURACY IS THE MOST IMPORTANT. If the instruction is NOT correctly followed, accuracy MUST be low (0-2). Even if the video looks realistic, incorrect instruction → low score.2. DO NOT BE FOOLED BY VISUAL QUALITY. Realism is not equal to correctness. Pretty but wrong → low score.3. PRESERVATION RULE: Only evaluate regions NOT related to the instruction.4. CONSISTENCY RULE: Check BOTH temporal and spatial stability across the entire video.5. BE STRICT: Most real-world outputs are imperfect → avoid giving 5 unless nearly perfect.

Spatial Track: 1) Accuracy (CRITICAL): Attribute accuracy (color, texture, size), Subject accuracy (add/remove/replace), Global accuracy (scene-level transformation). Reasoning: spatial relationships (position, layout, geometry) must be correct.2) Preservation: Unedited objects/regions must remain unchanged.3) Realism: Physical realism (lighting, shadow), Subject realism, Environment realism.4) Consistency: Temporal consistency (no flicker), Spatial consistency (structure stable).

Temporal Track: 1) Accuracy (CRITICAL): Action correctness, Attribute correctness (speed, freq), Timing correctness (order, delay), Camera control. Reasoning: temporal evolution must fol-

low real-world logic.2) Preservation: Unrelated temporal content must remain unchanged.3) Realism: Motion realism (natural, smooth, no artifacts).4) Consistency: Temporal stability (no drift).

Audio Track: 1) Accuracy (CRITICAL): Sound correctness, Timing correctness. Reasoning: audio must match event semantics.2) Preservation: Original audio must remain when not edited.3) Realism: Audio realism (no distortion), Scene consistency.4) Consistency: Audio-visual sync (lip sync).

Reference-based Track: 1) Accuracy (CRITICAL): All spatial/temporal edits must follow instruction. MUST align with the provided reference (image or description). Reference alignment is REQUIRED for a high score.2) Preservation: Non-edited content should remain unchanged.3) Realism: Physical plausibility and visual fidelity.4) Consistency: Temporal and spatial consistency.

Reasoning Track: 1) Accuracy (CRITICAL): Explicit instruction following + Implicit reasoning correctness (Temporal/Causal/Spatial/Logical evolution). IMPORTANT: If the underlying reasoning is wrong (e.g., cause-effect logic fails), accuracy must be LOW even if some edit exists.2) Preservation: Original visual content should be preserved unless required to change.3) Realism: Physical, subject, and environment realism.4) Consistency: Temporal stability and structural integrity.

G Visualization Results

We provide much more visualization results of models evaluated on our `OmniEdit-Bench` as shown in Fig. 6, Fig. 7 and Fig. 8.

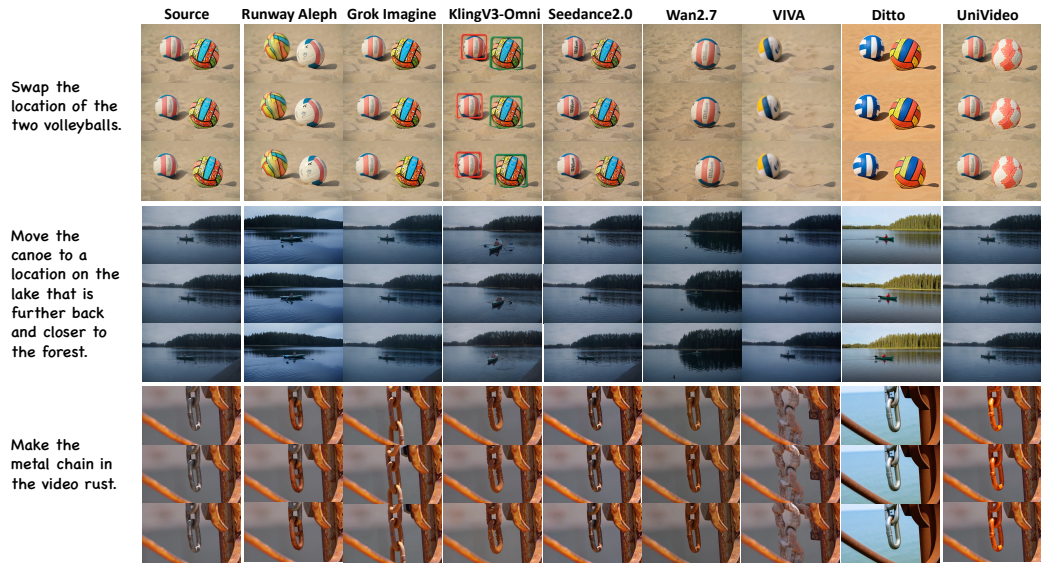

Figure 6

H Limitations

Despite its comprehensive design, `OmniEdit-Bench` still has several limitations. The benchmark is constrained by the scale and diversity of available data, particularly for complex temporal and audio scenarios. While VLM-based evaluation provides a scalable alternative to human annotation, it may not fully capture subtle subjective preferences or nuanced editing quality in borderline cases. In addition, the current designs of the reference-conditioned and reasoning tracks only cover a subset of possible settings, leaving more complex cross-modal dependencies and long-horizon reasoning scenarios insufficiently explored.

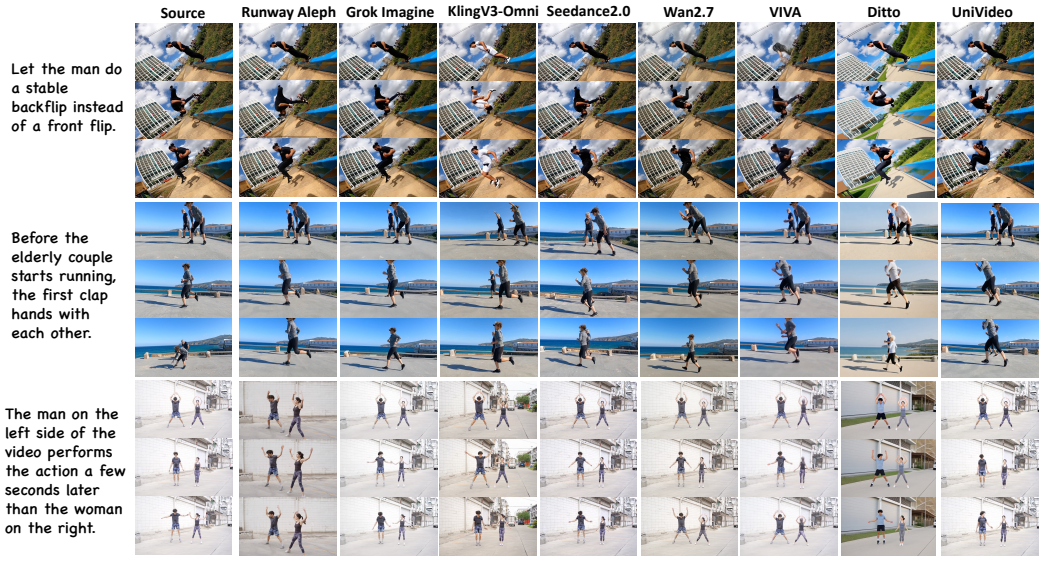

Figure 7

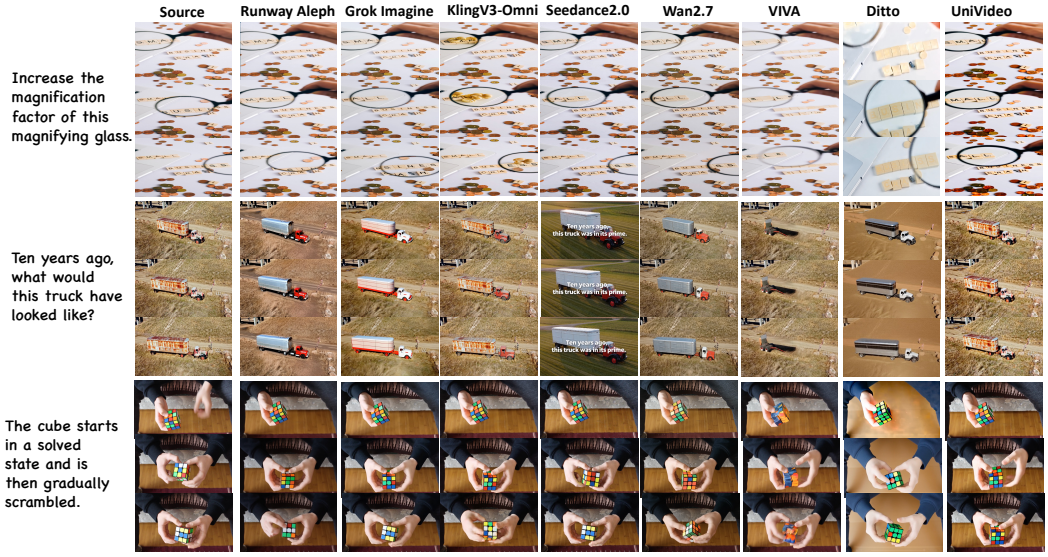

Figure 8

I Potential Impact

`OmniEdit-Bench` provides a comprehensive benchmark for evaluating instruction-based video editing, which can facilitate the development of more controllable and reliable video editing systems. Such advances have the potential to benefit a wide range of applications, including content creation, film production, education, and accessibility, by lowering the barrier for generating and editing video content through natural language.

At the same time, improved video editing capabilities may raise concerns about misuse, particularly in generating misleading or manipulated media. The availability of more powerful editing tools could exacerbate issues such as misinformation or synthetic content abuse. By providing a standardized and interpretable evaluation framework, our work aims to promote more responsible development and assessment of video editing systems. Future work should further explore safeguards, including detection and attribution mechanisms, to mitigate potential risks.