Kiwi-Edit: Versatile Video Editing via Instruction and Reference Guidance

Yiqi Lin Guoqiang Liang Ziyun Zeng Zechen Bai Yanzhe Chen Mike Zheng Shou ✉

Show Lab, National University of Singapore

Project Page: https://showlab.github.io/Kiwi-Edit

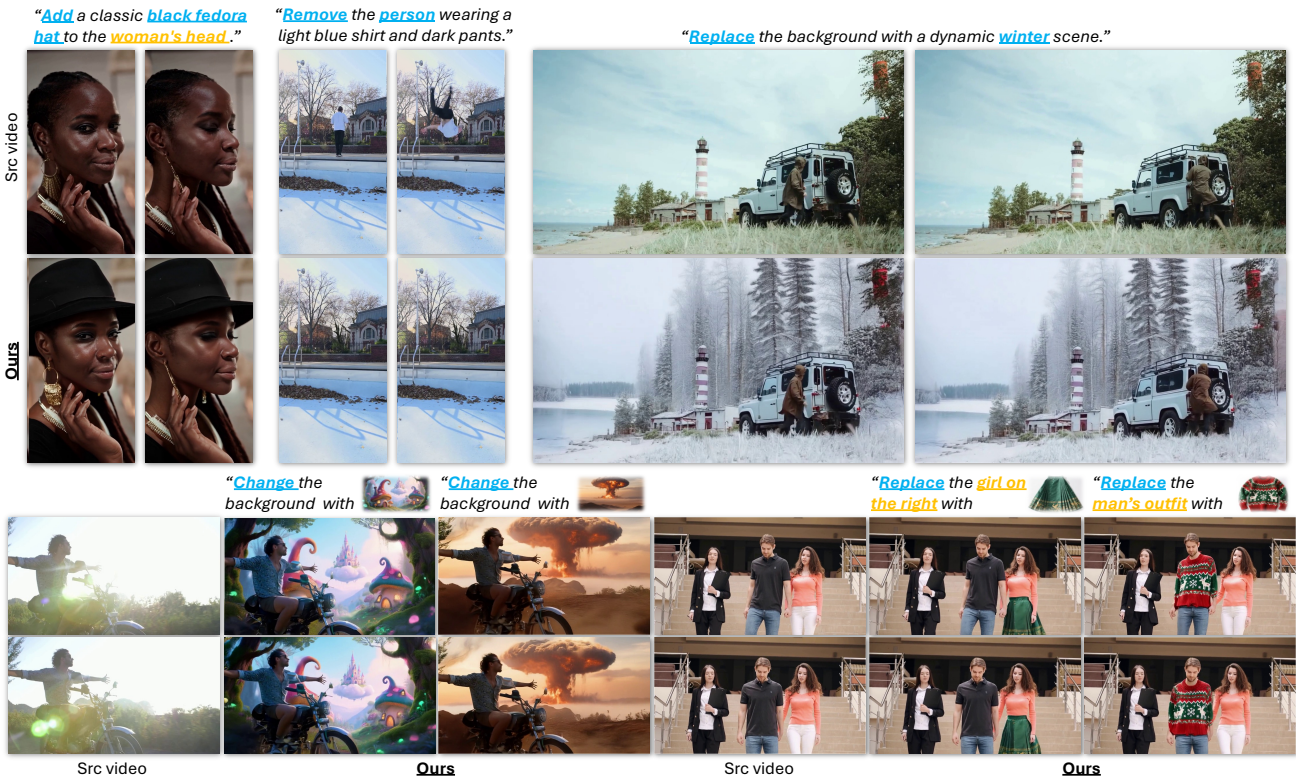


Figure 1. This teaser illustrates a selection of video editing tasks, including both instruction-only and instruction-reference scenarios, highlighting the superior editing capabilities of RefVIE.

Abstract

Instruction-based video editing has witnessed rapid progress, yet current methods often struggle with precise visual control, as natural language is inherently limited in describing complex visual nuances. Although reference-guided editing offers a robust solution, its potential is currently bottlenecked by the scarcity of high-quality paired training data. To bridge this gap, we introduce a scalable data generation pipeline that transforms existing video editing pairs into high-fidelity training quadruplets, leveraging image generative models to create synthesized reference scaffolds. Using this pipeline, we construct RefVIE, a large-scale dataset tailored for instruction-reference-following tasks, and establish RefVIE-Bench for comprehensive evaluation. Furthermore, we propose a unified editing architecture, Kiwi-Edit, that synergizes learnable queries and latent visual features for reference semantic guidance. Our model achieves significant gains in instruction following and reference fidelity via a progressive multi-stage training curriculum. Extensive experiments demonstrate that our data and architecture establish a new

. Correspondence to: Mike Zheng Shou <mike.zheng.shou@gmail.com>.

Preprint. May 14, 2026.

state-of-the-art in controllable video editing. *All datasets, models, and code is released at* https://github.com/showlab/Kiwi-Edit.

1. Introduction

The customization of video content across social media, entertainment, and advertising has fueled an unprecedented demand for accessible video editing tools. Recent advances in instruction-based video editing (Cheng et al., 2023; Bai et al., 2025a; Zi et al., 2025; Wu et al., 2025b; He et al., 2025) have demonstrated remarkable progress, enabling users to modify video content through natural language commands. These methods leverage powerful video diffusion models (Wan et al., 2025; Kong et al., 2024) to execute diverse editing operations, ranging from local object modifications to global style transfers, while preserving the content and temporal coherence across frames.

Despite this progress, a critical limitation persists: the reliance on text-only instructions. Natural language is inherently ambiguous when describing precise visual details such as specific textures, exact object identities, or nuanced stylistic characteristics. Users frequently desire to convey editing intent through visual examples, such as "replace the car with *this* sports car" or "apply the style of *this* painting," yet text-only models fundamentally struggle to accomplish such tasks. Reference-guided video editing, which conditions generation on both textual instructions and visual references, offers a natural solution to this challenge.

However, the development of reference-guided video editing is severely constrained by data scarcity. Training such models requires high-quality quadruplets comprising source videos, editing instructions, reference images, and target videos, a format that existing datasets do not provide at scale. As summarized in Table 1, while several large-scale instruction-based video editing datasets exist (Wu et al., 2025b; Zi et al., 2025; Bai et al., 2025a; Zhang et al., 2025; He et al., 2025), *none offer reference image*. The few works that explore reference-guided editing (Mou et al., 2025; Team et al., 2025) rely on proprietary data that remain inaccessible to the wider research community. This bottleneck fundamentally impedes the field.

To address this challenge, we curate RefVIE, a large-scale dataset for instruction-reference guided video editing. Our key insight is that powerful pre-trained image generation models can serve as high-fidelity reference synthesizers, enabling scalable data construction without expensive manual annotation. Starting from existing instruction-based video editing datasets (Bai et al., 2025a; Zhang et al., 2025; He et al., 2025) that provide source-target video pairs, we design an automated pipeline to synthesize the missing reference images. Specifically, we leverage vision-language models to ground editing regions in video frames, followed by state-of-the-art image editors to generate reference images that capture the visual essence of the intended edit. Through rigorous quality filtering and de-duplication, we construct a dataset of **477K** high-quality quadruplets from an initial pool of 3.7M samples. To the best of our knowledge, **RefVIE is the first large-scale, open-source resource for instruction-reference guided video editing.**

Table 1. A summary of existing large scale instruction and reference guided video editing datasets.

Dataset	Open-Source	Instr. Edit	Ref. Image	Num.
InsViE (Wu et al., 2025b)	✓	✓	✗	1M
Señorita (Zi et al., 2025)	✓	✓	✗	2M
Ditto (Bai et al., 2025a)	✓	✓	✗	1M
ReCo (Zhang et al., 2025)	✓	✓	✗	500K
OpenVE (He et al., 2025)	✓	✓	✗	3M
InsturctX (Mou et al., 2025)	✗	✓	✓	236K
Kling-Omni (Team et al., 2025)	✗	✓	✓	—
RefVIE (Ours)	✓	✓	✓	477K

Building upon this data foundation, we develop **Kiwi-Edit**, a unified video editing framework that effectively integrates multimodal conditions. Our architecture couples a frozen Multimodal Large Language Model (MLLM) with a Diffusion Transformer (DiT), where the MLLM processes interleaved sequences of source video frames, textual instructions, and reference images. We employ a dual-connector mechanism: a *Query Connector* that projects learnable query tokens to distill editing intent, and a *Latent Connector* that extracts visual features from reference images. These connectors produce unified context tokens that guide the DiT via cross-attention. To preserve source video structure while enabling flexible reference-guided editing, we introduce a hybrid latent injection strategy: source video features are added element-wise with a timestep-dependent scalar for structural preservation, while reference image features are concatenated to the input sequence for fine-grained texture transfer. A three-stage curriculum training strategy ensures stable convergence: MLLM-DiT alignment, instructional tuning, and reference-guided fine-tuning. Furthermore, to enable rigorous evaluation, we establish RefVIE-Bench, a benchmark of 100 manually verified samples designed to assess reference adherence, instruction compliance, and temporal consistency.

In summary, our contributions are threefold: First, we curate RefVIE, a large-scale dataset of 477K high-quality quadruplets for instruction-reference guided video editing, covering local editing and background replacement. Second, we introduce RefVIE-Bench, a comprehensive benchmark specifically designed to evaluate reference similarity, instruction accuracy, and temporal consistency. Lastly, we present a unified video editing model that achieves state-of-the-art performance on both instruction-only and reference-guided tasks through a novel MLLM-DiT architecture design and multi-stage training curriculum.

2. Related Work

2.1. Instruction-based Video Editing

Given the challenges in training robust text-to-video models from scratch, early prevalent approaches (Wu et al., 2023; Cong et al., 2023; Geyer et al., 2023; Kara et al., 2024; Ku et al., 2024; Qi et al., 2023) leverage the rich generative priors of pre-trained text-to-image (T2I) models. These methods typically employ fine-tuning or inversion techniques (Song et al., 2020) to achieve instruction-guided editing. However, T2I-based methods often suffer from limited temporal consistency and inversion artifacts, particularly under complex motion or occlusions. Consequently, with the advent of open-source video diffusion models (Hong et al., 2022; Kong et al., 2024; Wan et al., 2025), recent research has shifted towards utilizing these native video backbones to ensure better motion fidelity. To support instruction-based training, InsV2V (Cheng et al., 2023) pioneers the field by utilizing InstructPix2Pix (Brooks et al., 2023) to synthesize paired training data. Addressing the fundamental challenge of data scarcity, subsequent efforts (Bai et al., 2025a; Zi et al., 2025; Wu et al., 2025b) focus on constructing large-scale synthesized datasets. For instance, Senorita-2M (Zi et al., 2025) collects editing pairs via a mixture-of-experts pipeline, while Ditto (Bai et al., 2025a) leverages edited key-frames and depth maps to generate edited videos. More recently, to enhance instruction following and semantic understanding, Omni-Video (Tan et al., 2025) and OpenVE-Edit (He et al., 2025) integrate Vision-Language Models (VLMs) into the editing framework.

2.2. Reference-Guided Video Editing and Dataset

Relying solely on natural language prompts often fails to capture the nuances of visual imagination. Text is inherently limited in describing precise spatial relationships, specific visual references, and temporal dynamics, creating a gap between user intent and model output. To address this, recent works (Mou et al., 2025; Wei et al., 2025; Team et al., 2025) have introduced reference images alongside textual instructions to enable more precise video editing. Specifically, methods like InstructX (Mou et al., 2025) and Kling-Omni (Team et al., 2025) feed multi-modal inputs into MLLMs to extract unified representations for the generation module. However, despite their impressive performance, these approaches rely heavily on proprietary in-house models for data generation and require extensive manual verification to curate high-quality reference data. A summary of represented video editing datasets is presented in Table 1. In this work, we aim to democratize this capability by introducing **RefVIE**, the first large-scale open-source dataset tailored for instruction-reference guided video editing. By providing a rigorously filtered and curated dataset, we offer a critical resource to the research community, enabling the development of comprehensive editing models that effectively transcend the limitations of text-only control.

3. RefVIE Dataset and Benchmark

3.1. Scalable Data Generation Pipeline

A primary bottleneck in advancing reference-guided video editing is the scarcity of high-quality training quadruplets: $(V_{src}, T_{inst}, I_{ref}, V_{tgt})$. Manual curation of such 4-tuple data is prohibitively expensive. To address this, we propose a scalable, automated pipeline to synthesize reference images from existing instruction-based video editing pairs, effectively augmenting standard triplets $(V_{src}, T_{inst}, V_{tgt})$ into the required quadruplets. As illustrated in Figure 3, our pipeline processes a massive pool of 3.7M raw samples from publicly available datasets through four distinct stages.

Stage 1. Source Aggregation and Filtering. We initialize our data pool by aggregating three open-source instructional video editing datasets: Ditto-1M (Bai et al., 2025a), ReCo (Zhang et al., 2025), and OpenVE-3M (He et al., 2025). To ensure high training quality, we filter these samples using EditScore (Luo et al., 2025). Based on a human-pivoted pilot study, we discard samples with an EditScore lower than 6 for text-guided instruction tuning. For reference-guided generation specifically, we apply a stricter threshold (EditScore > 8) and explicitly select tasks categorized as *Local Modification* or *Background Replacement*, as these benefit most from visual referencing.

Stage 2. Grounding and Segmentation. Precise spatial localization is critical for generating consistent references. We employ Qwen3-VL-32B (Bai et al., 2025b) to interpret the editing instruction and ground the region of interest in the first frame of the *target* video, as the target video contains the desired editing result from which we extract the reference. For *Background Change*, the model grounds the foreground object so that it can be removed in the next stage, leaving only the new background as the reference. For *Local Editing*, the model grounds the edited object so that it can be extracted as the reference. These coarse bounding box coordinates are then refined by SAM3 (Carion et al., 2025) to produce pixel-perfect segmentation masks. Pairs that fail the grounding or segmentation checks are discarded.

Stage 3. Reference Image Synthesis. Leveraging the segmented regions, we synthesize reference images using Qwen-Image-Edit-2511 (Wu et al., 2025a), as depicted in Figure 2. For background tasks, we extract and remove the foreground object, then inpaint the region to produce a clean background image that serves as the reference. For local edits, we extract the target object and place it on a clean background with minimal surrounding space, creating a tightly cropped reference that highlights the edited object's appearance. To ensure data robustness, we also filter out generated images with extreme aspect ratios or resolution.

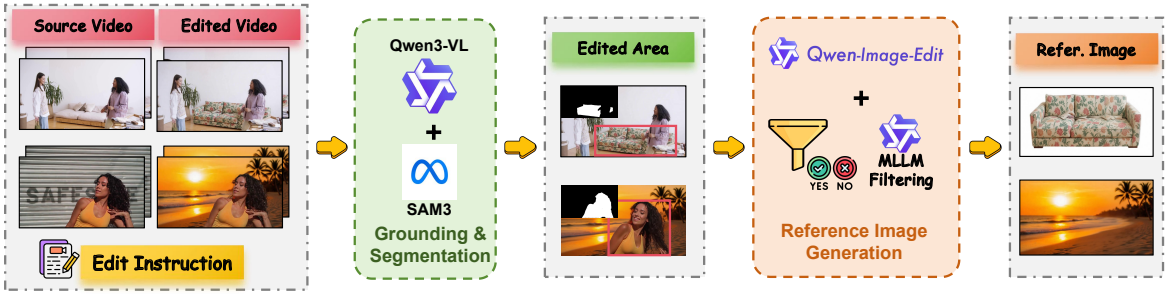


Figure 2. Workflow of the reference image synthesis pipeline. We first ground the editing region in the target video frame using specialized grounding and segmentation models. Subsequently, we leverage a specialized image editing model to synthesize a high-quality reference image that maintains identity consistency with the instruction.

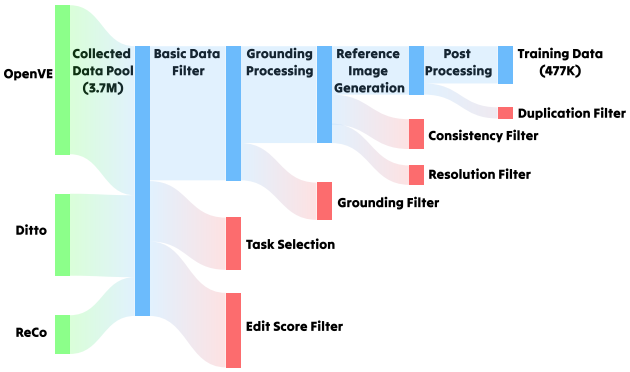
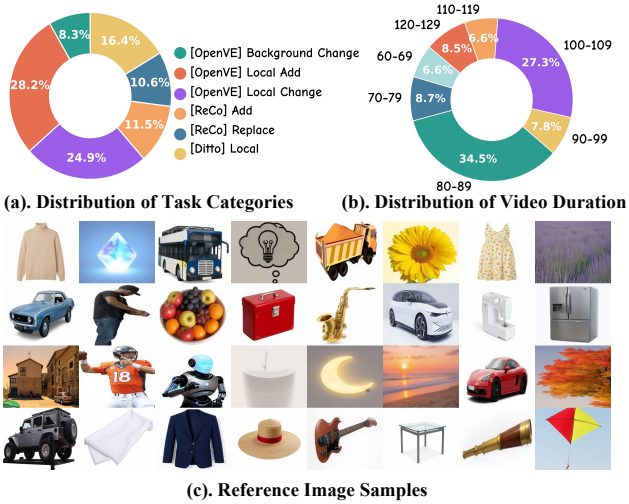


Figure 3. Pipeline of RefVIE curation. We process 3.7M raw samples through four stages: source aggregation and filtering, grounding and segmentation, reference image synthesis, and quality control, yielding 477K high-quality quadruplets.



(a). Distribution of Task Categories (b). Distribution of Video Duration

(c). Reference Image Samples

Figure 4. RefVIE statistics and sample visualization. (a) Distribution of editing task types. (b) Distribution of video durations. (c) Example reference images for different editing categories.

Stage 4. Quality Control and Post-Processing. In the final stage, we enforce semantic alignment by using an MLLM to verify that the synthesized reference image is consistent with the edited content in the target video, filtering out low-fidelity generations. Additionally, to prevent data leakage and redundancy, we extract CLIP (Radford et al., 2021) features from the reference images and perform global deduplication. This rigorous pipeline distills the initial 3.7M pool down to a high-quality subset of 477K instruction-reference-video quadruplets. The detailed breakdown across filtering stages is visualized in Figure 3.

3.2. Dataset Statistics

Our resulting dataset, **RefVIE**, is the largest open-source collection for reference-guided video editing, bridging the gap between academic resources and commercial capabilities. Figure 4 summarizes the dataset statistics. As shown in Figure 4(a), the task distribution is well-balanced across local object addition, replacement, and background changes, ensuring that trained models generalize across different editing scenarios rather than overfitting to a single task type. Figure 4(b) illustrates the video duration distribution. Most clips contain 80 to 110 frames, providing sufficient temporal context for models to learn long-range motion consistency and handle complex object movements. Figure 4(c) presents example reference images, demonstrating the diversity of visual content including objects, textures, and backgrounds. More visualization cases are provided in the supplementary.

3.3. Benchmark and Evaluation

RefVIE-Bench. Existing benchmarks predominantly focus on text-video alignment, neglecting the critical dimension of visual reference adherence. To rigorously evaluate this capability, we establish RefVIE-Bench, comprising 110 manually verified triplets $(V_{src}, I_{ref}, T_{ins})$. Unlike our scalable synthetic training data, these benchmark samples undergo a rigorous three-stage manual verification process to ensure high quality and diversity. The benchmark evaluates two specific capabilities: *Subject Reference* (70 samples) and *Background Replacement* (40 samples). The larger allocation to object modification reflects its broader scope, encompassing diverse object categories such as vehicles, animals, clothing, and furniture, each with distinct visual attributes. In contrast, Background replacement represents a more unified task category focused on environment changes

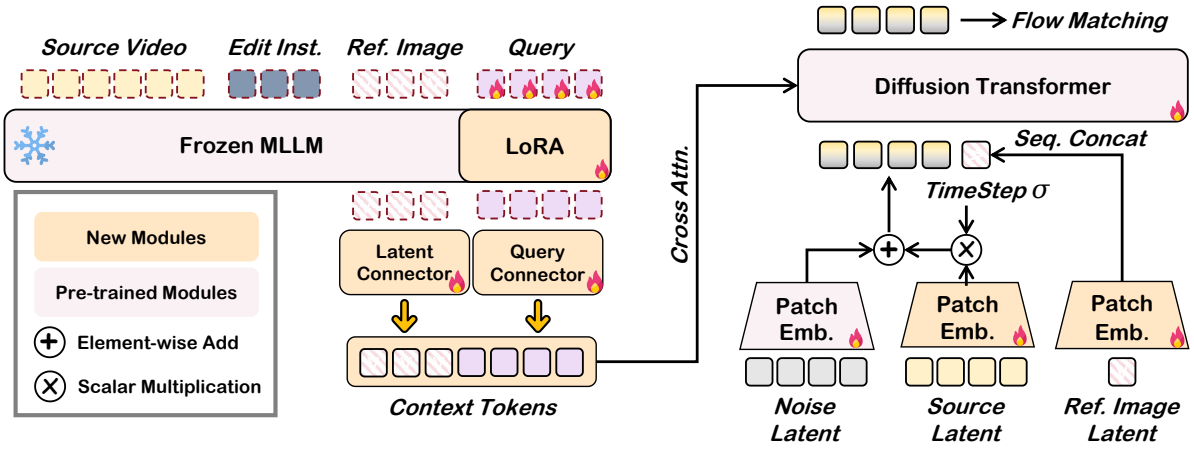


Figure 5. Overview of our unified editing framework. We integrate a frozen MLLM (Qwen2.5-VL-3B) to encode multimodal instructions, injecting semantic conditions into the pre-trained Diffusion Transformer (Wan2.2-TI2V-5B) via dual learnable projectors for query and reference latents. To preserve consistency of source video, we employ a hybrid injection strategy within the DiT: source video features are added element-wise, while reference image features are concatenated to the input sequence.

while preserving foreground dynamics.

Evaluation Metrics. Traditional metrics like CLIP score only capture high-level semantic similarity, while FID measures distribution-level statistics rather than per-sample quality. Neither can assess whether specific textures are preserved or whether editing instructions are correctly followed. To address this, we employ a state-of-the-art MLLM (Gemini3 (Comanici et al., 2025)) as an automated judge, which evaluates each result on a scale of 1 to 5 across three dimensions. For subject reference, we evaluate *Identity Consistency*, *Temporal Fidelity*, and *Physical Integration* (e.g., tracking, shadows), while for background replacement, we assess *Reference Fidelity*, *Matting Quality*, and *Visual Harmony* (e.g., perspective, lighting). To ensure logical robustness, we enforce a hierarchical constraint where temporal and physical scores are capped by the primary identity score, preventing the model from assigning high ratings to edits that are temporally stable but semantically incorrect.

4. Methodology

4.1. Architecture Design

As illustrated in Figure 5, our framework consists of two main components: a Multimodal Large Language Model (MLLM) for semantic understanding and a Diffusion Transformer (DiT) for video generation. The MLLM encodes multimodal inputs (source video, instruction, and optional reference image) into conditioning signals, which guide the DiT to generate the edited video.

Semantic Conditioning via MLLM. We utilize Qwen2.5-VL-3B (Bai et al., 2025c) as the MLLM backbone. The base model weights are frozen, and we inject lightweight Low-Rank Adaptation (Hu et al., 2022) (LoRA) modules to adapt it to the video editing domain without compromising pre-trained knowledge. The MLLM processes an interleaved sequence comprising the source video frames, textual editing instructions, and optional reference images. From its output, we extract conditioning features through two specialized pathways: *Instructional Queries:* We utilize a set of learnable query tokens (256 for image tasks, 512 for video editing, 768 for reference task) to distill the editing intent (e.g., "turn the sky red"). These are projected via a *Query Connector* (MLP) to align with the DiT's dimension. *Reference Latents:* For tasks requiring specific visual guidance, we extract visual tokens corresponding to the reference image. These are projected via a separate *Latent Connector*. The outputs of these connectors are concatenated to form a unified sequence of *Context Tokens*, which serve as the key/value pairs for the DiT's cross-attention layers, guiding the semantic content of the generation.

Structural Conditioning via Latent Injection. While MLLM context provides semantic guidance, preserving the precise structural layout of the source video requires a more direct signal. We identify that standard cross-attention is insufficient for fine-grained spatial preservation. Therefore, we introduce a hybrid injection strategy.

Source Video Control (Element-wise Injection): To preserve the spatial-temporal structure, we encode the source frames into latent space using the VAE. These latents are processed by a zero-initialized `PatchEmbed` layer. As depicted in Figure 5, rather than concatenating these features (which we found leads to training instability), we add them element-wise to the noisy latent $\mathbf{z}_t$. Crucially, this addition is modulated by a time-dependent factor $\sigma(t)$:

$$\begin{aligned}\mathbf{z}'_t = &\texttt{PatchEmbed}(\mathbf{z}_t) \\ &+ \sigma(t) \cdot \texttt{PatchEmbed}_{src}(\text{VAE}(\mathbf{x}_{src})).\end{aligned} \tag{1}$$

Table 2. OpenVE-Bench results evaluated on Gemini-2.5-Pro. Orange rows denote closed-source models and are excluded when selecting the best and second-best scores; gray rows denote open-source baselines and blue rows denote our variants. Bold and underlined scores indicate the best and second-best results among non-closed-source methods, respectively.

Method	#Params. (DiT)	#Reso.	Overall↑	Global Style ↑	Background Change ↑	Local Change ↑	Local Remove ↑	Local Add ↑
Runway Aleph	-	1280×720	3.49	3.72	2.62	4.18	4.16	2.78
VACE (Jiang et al., 2025)	14B	1280×720	1.57	1.49	1.55	2.07	1.46	1.26
OmniVideo (Tan et al., 2025)	1.3B	640×352	1.19	1.11	1.18	1.14	1.14	1.36
InsViE (Wu et al., 2025b)	2B	720×480	1.45	2.20	1.06	1.48	1.36	1.17
Lucy-Edit (Team, 2025)	5B	1280×704	2.22	2.27	1.57	3.20	1.75	2.30
ICVE (Liao et al., 2025)	13B	384×240	2.18	2.22	1.62	2.57	2.51	1.97
DITTO (Bai et al., 2025a)	14B	832×480	2.13	**4.01**	1.68	2.03	1.53	1.41
OpenVE-Edit (He et al., 2025)	5B	1280×704	2.50	3.16	2.36	2.98	1.85	2.15
ReCo (Zhang et al., 2025)	2.1B	832×480	2.80	<u>3.96</u>	1.92	3.70	2.24	2.17
UniVideo (Wei et al., 2025)	14B	854×480	<u>3.02</u>	3.67	2.51	3.79	**2.73**	<u>2.40</u>
Ours (Stage-2 Instruct-Only)	5B	720 × 480	2.92	3.54	<u>2.59</u>	3.80	2.55	2.12
Ours (Stage-2 Instruct-Only)	5B	1280 × 704	2.98	3.54	2.57	<u>3.84</u>	<u>2.71</u>	2.25
Ours (Stage-3 Instruct-Reference)	5B	1280 × 704	**3.11**	3.72	**2.67**	**3.91**	2.69	**2.55**

Here $\sigma(t)$ is the time-dependent factor sampled from noise schedule. Our ablation studies confirm that this factor is critical; removing it causes the model to ignore the detail source structure, while replacing addition with channel concatenation degrades editability (see Table 4).

Reference Image Control (Sequence Concatenation): Conversely, to enforce high fidelity to a reference object, we treat the reference image as an extension of the visual context. The reference image $\mathbf{x}_{ref}$ is patch-embedded and *concatenated* to the input sequence of the DiT. This effectively extends the spatial-temporal attention window, allowing the model to "copy" texture details from the reference directly.

Training Objective. We employ Flow Matching (Lipman et al., 2022) as the training objective to minimize the mean squared error between the predicted velocity field $\mathbf{v}_\theta$ and the ground-truth drift:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1,\mathbf{c}}\left[||\mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{c}) - (\mathbf{z}_1 - \mathbf{z}_0)||^2\right] \quad (2)$$

where $\mathbf{z}_1$ is the target video latent, $\mathbf{z}_0$ is standard Gaussian noise, and $\mathbf{c}$ represents the multimodal conditioning signals.

4.2. Training Curriculum

Training a billion-parameter video generation model with multimodal conditions is computationally intensive and prone to optimization difficulties. To ensure stable convergence and effective alignment, we adopt a multi-stage progressive training curriculum as follows:

Stage 1. MLLM-DiT Alignment. In the initial phase, we freeze both the MLLM base weights and DiT backbone. We train only the bridge components: the LoRA adapters, the Query/Latent Connectors, and the learnable query tokens. This stage uses text-based editing triplets $(V_{src}, T_{inst}, V_{tgt})$ and focuses on establishing a semantic mapping, ensuring the Connectors can translate MLLM representations into a format interpretable by the DiT's cross-attention blocks. In this stage, the training data consists exclusively of high-quality image editing tasks sourced from GPT-Image-Edit (Wang et al., 2025) and NHR-Edit (Kuprashevich et al., 2025), which provide a efficient way to align semantic space (Pan et al., 2025).

Stage 2. Instructional Tuning. We subsequently unfreeze the DiT layers to enable joint optimization. The model continues training on text-based editing triplets $(V_{src}, T_{inst}, V_{tgt})$ from large-scale instructional image and video editing datasets. The training data for this stage combines the image datasets from Stage 1 with our curated instruction video editing subset (filtered with EditScore ≥ 6) as detailed in Section 3.1. This stage is crucial for learning general editing primitives (e.g., object removal, style transfer). To improve efficiency, we employ a resolution curriculum, initializing training with low-resolution clips (480p) and progressively scaling to higher resolutions (720p).

Stage 3. Reference-Guided Fine-tuning. In the final stage, we introduce our curated RefVIE dataset to unlock precise visual control. We train on a mixture of instruction editing data from Stage 2 and new reference-guided quadruplets $(V_{src}, T_{inst}, I_{ref}, V_{tgt})$. This stage refines the model's ability to utilize the reference tokens for fine-grained textural transfer, ensuring the generated content aligns with user-provided visual examples. During all training stages, we set the maximum number of frames sampled from video to 81.

5. Experiments

5.1. Implementation Details

For image data, we utilize triplets (instruction, source image, edited image) dataset from NHR-Edit (Kuprashevich et al., 2025) and GPT-Image-Edit (Wang et al., 2025). We use the

Table 3. Comparison on RefVIE-Bench with *Gemini-3-Flash* as Judge Model.

Model	#Params.(DiT)	Subject Reference				Background Reference				Overall↑
		Identity Consist.	Temporal Consist.	Physical Consist.	Subj. Avg.	Refer. Sim.	Matting Quality	Video Quality	BG Avg.	
Runway Aleph	-	3.79	3.65	3.58	3.67	3.33	2.81	2.58	2.91	3.29
Kling-O1 (Team et al., 2025)	-	4.75	4.66	4.60	4.67	3.95	3.21	2.75	3.30	3.99
UniVideo (Wei et al., 2025)	14B	4.17	**3.79**	**3.59**	**3.85**	3.13	2.55	2.28	2.65	**3.44**
ReCo-Ref (Zhang et al., 2025)	2.1B	3.65	2.95	2.70	3.10	3.35	**2.65**	2.38	**2.79**	3.00
Ours (Stage-3 Insturuct-Reference)	5B	3.51	2.96	2.91	3.13	**3.40**	2.58	**2.40**	**2.79**	2.96
Ours (Stage-3 Reference. only)	5B	**4.28**	3.61	3.55	3.81	3.33	2.35	2.08	2.58	3.40

Figure 6. Qualitative results in OpenVE-Bench and VIE-Bench. Please zoom in for more details.

pretrained Wan-TI2V-5B as the DiT backbone and finetune it on open source dataset and our propose reference data for 12K steps. We set the learning rate to 2 × 10-5, with a global batch size of 128 with gradient accumulation. In stage 2, we sample image and instruction video data at 1:1 ratio and first train on 360K pixels and then 960K pixels for 10K steps. In stage 3, we sample image and instruction video data and video with reference image at 2:1:1 ratio for 10K steps. More details are in supplementary materials.

5.2. Main Results

Instruction Editing. We compare our model with existing state-of-the-art open-source models on OpenVE-Benchmark (He et al., 2025), including VACE (Jiang et al., 2025), OmniVideo (Tan et al., 2025), InsViE (Wu et al., 2025b), ICVE (Liao et al., 2025), Lucy-Edit (Team, 2025), and DITTO (Bai et al., 2025a), as well as the closed-source model Runway Aleph. The evaluation setting follows the original setting report in the paper. As shown in Table 1 , our method outperforms all open-source baselines with an Overall score of **3.11**, significantly surpassing the previous best, OpenVE-Edit 2.50. Notably, our model excels in *Background Change*, achieving a score of 2.67 which surpasses even the proprietary Runway Aleph 2.62. Additionally, increasing inference resolution to 1280 × 704 and applying training curriculum yields consistent performance gains across all metrics.

Instruction and Reference Guided Editing. We compare our method against leading commercial video generation models, specifically Runway Aleph and Kling-O1 (Team et al., 2025). As summarized in Table 3, our model trained on the curated *RefVIE* achieves an overall score of **3.40**. Notably, our model demonstrates competitive performance in *Identity Consistency* 4.28 and *Reference Similarity* 3.40. While the proprietary Kling-O1 model achieves higher absolute scores, likely attributable to its significantly larger parameter count and closed-source training corpus, we setup a state-of-the-art baseline for open-source reference guided video editing.

5.3. Qualitative Results

Figure 6 and Figure 7 present visual comparisons across diverse editing tasks. Our model exhibits superior instruction following and reference preservation. *Instruction Following:* Our model accurately captures visual semantics from the source and reference. For example, it correctly localizes the hat (Figure 6, Col. 3) and the table (Figure 7, Row 2). *Reference Consistency:* As shown by the red bounding boxes in Figure 7 (Row 3), our method maintains high subject consistency even during drastic background style changes. *More qualitative results are provided in the appendix and supplementary material.*

5.4. Ablation Studies

Condition Design. We conduct an ablation study on the model structure of source video input conditioning, compar-

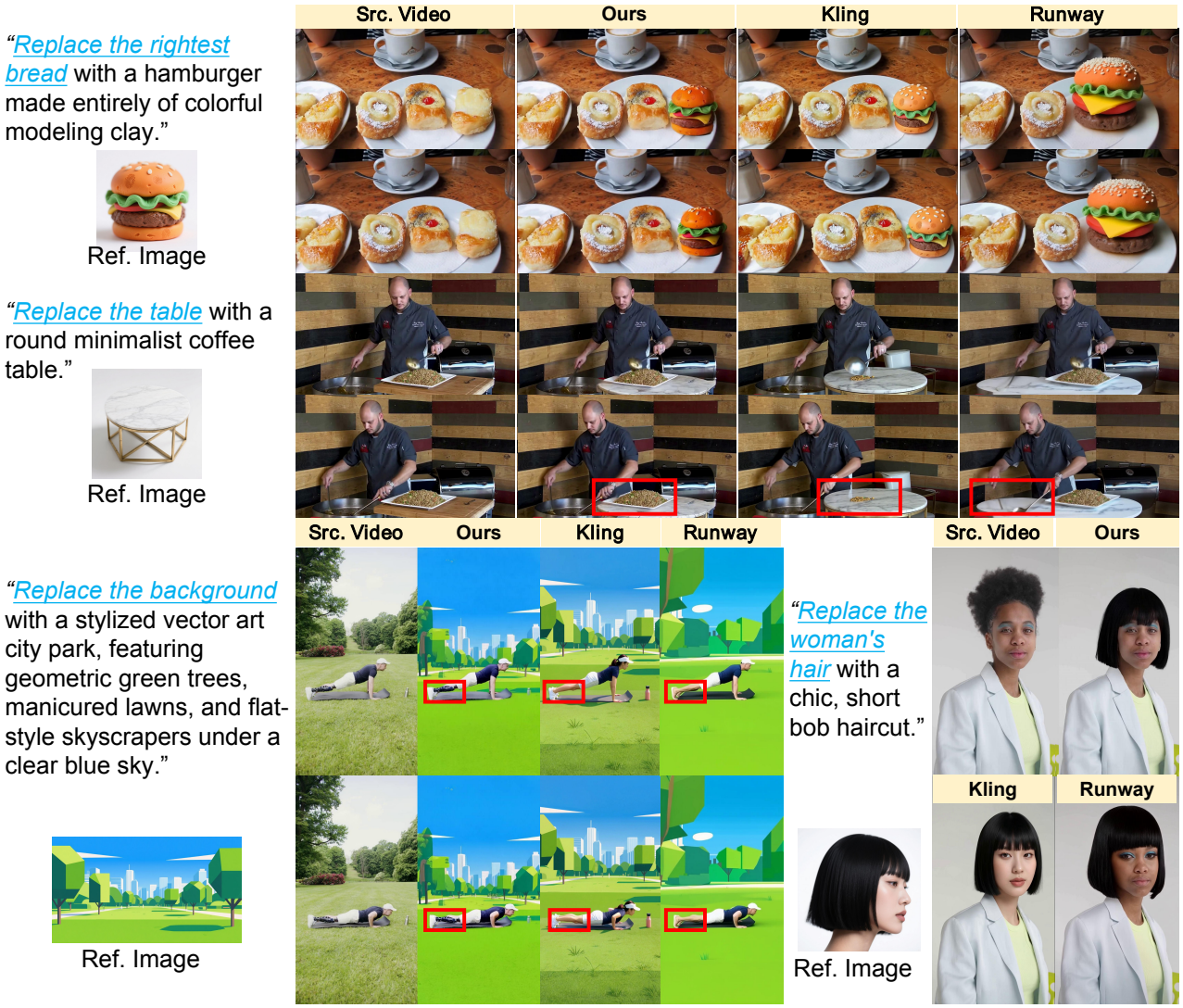


Figure 7. Qualitative results in our proposed RefVIE-Bench. Please zoom in for more details.

Table 4. Ablation on Conditional Design and using Image data.

Method	Score@Remove ↑	Score@Style ↑
Add w/ time-dependent factor.	**2.63**	**4.07**
Add wo/ time-dependent factor.	2.58	4.05
Add (Shared Patch Embedding)	1.01	1.00
Channel Concat	2.08	3.82
Baseline (default)	**2.84**	3.98
w/o Alignment	1.47	3.01
w/o Image Co-train	2.58	**4.07**

ing channel concatenation with feature addition strategies. As shown in Table 4, Channel Concat performs poorly, while sharing patch embeddings significantly degrades results to 1.01, confirming the need for independent feature extraction. The 'Add w/ time-dependent factor' configuration proves most effective, outperforming the baseline in instruction tasks including remove (2.63) and add (2.01).

Training Curriculum. We perform an ablation study to validate the necessity of our progressive training stages, with results summarized in Table 4. First, skipping the *Alignment* stage leads to a catastrophic performance drop, confirming that establishing a coarse semantic mapping between the MLLM and DiT is a prerequisite for effective instruction following. Second, excluding *Image Co-training* degrades performance on structural tasks (Removal score 2.84 → 2.58). This indicates that while video-only training can achieve high style scores 4.07, it lacks the fine-grained spatial supervision provided by image editing datasets, which is essential for complex local manipulations.

Reference Condition Design. To validate the effectiveness of our dual-connector design, we analyze the impact of different condition injection pathways in the MLLM. As shown in Table 5, relying solely on learnable instructional queries yields a baseline score of 3.20. While queries effectively capture the high-level editing intent, they often struggle to preserve fine-grained visual details. By incorporating

Table 5. Ablation on architecture choice.

Query	Ref. Latent	**Score@Subject** ↑
✓	✗	3.20
✓	✓	3.30

the Reference Latent features via the Latent Connector, we explicitly inject dense semantic priors from the reference image into the context. This addition improves the score to 3.30, demonstrating that combining sparse instructional queries with dense visual latents is essential for achieving high-fidelity reference adherence.

6. Conclusion

This work addresses the critical scarcity of high-quality data for reference-guided video editing. We introduce a scalable pipeline to synthesize RefVIE, transforming existing video pairs into rich instruction-reference quadruplets, and establish RefVIE-Bench to standardize evaluation in this domain. Leveraging this data, our unified edit model **Kiwi-Edit** achieves state-of-the-art performance, effectively bridging the gap between user intent and video generation. We believe this data-centric approach lays a solid foundation for more controllable and accessible video content creation.

References

Bai, Q., Wang, Q., Ouyang, H., Yu, Y., Wang, H., Wang, W., Cheng, K. L., Ma, S., Zeng, Y., Liu, Z., et al. Scaling instruction-based video editing with a high-quality synthetic dataset. *arXiv preprint arXiv:2510.15742*, 2025a.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X.-H., Cheng, Z., Deng, L., Ding, W., Fang, R., Gao, C., et al. Qwen3-vl technical report. *ArXiv*, abs/2511.21631, 2025b. URL https://api.semanticscholar.org/CorpusID:283262018.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025c.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K. V., Khedr, H., Huang, A., et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Cheng, J., Xiao, T., and He, T. Consistent video-to-video transfer using synthetic dataset. *arXiv preprint arXiv:2311.00213*, 2023.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Cong, Y., Xu, M., Simon, C., Chen, S., Ren, J., Xie, Y., Perez-Rua, J.-M., Rosenhahn, B., Xiang, T., and He, S. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023.

Geyer, M., Bar-Tal, O., Bagon, S., and Dekel, T. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

He, H., Wang, J., Zhang, J., Xue, Z., Bu, X., Yang, Q., Wen, S., and Xie, L. Openve-3m: A large-scale high-quality dataset for instruction-guided video editing. *arXiv preprint arXiv:2512.07826*, 2025.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., and Liu, Y. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

Kara, O., Kurtkaya, B., Yesiltepe, H., Rehg, J. M., and Yanardag, P. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6507–6516, 2024.

Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Ku, M., Wei, C., Ren, W., Yang, H., and Chen, W. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

Kuprashevich, M., Alekseenko, G., Tolstykh, I., Fedorov, G., Suleimanov, B., Dokholyan, V., and Gordeev, A. Nohumansrequired: Autonomous high-quality image editing triplet mining. *arXiv preprint arXiv:2507.14119*, 2025.

Liao, X., Zeng, X., Song, Z., Fu, Z., Yu, G., and Lin, G. In-context learning with unpaired clips for instruction-based video editing. *arXiv preprint arXiv:2510.14648*, 2025.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Luo, X., Wang, J., Wu, C., Xiao, S., Jiang, X., Lian, D., Zhang, J., Liu, D., et al. Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. *arXiv preprint arXiv:2509.23909*, 2025.

Mou, C., Sun, Q., Wu, Y., Zhang, P., Li, X., Ye, F., Zhao, S., and He, Q. Instructx: Towards unified visual editing with mllm guidance. *arXiv preprint arXiv:2510.08485*, 2025.

Pan, X., Shukla, S. N., Singh, A., Zhao, Z., Mishra, S. K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., and Chen, Q. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Tan, Z., Yang, H., Qin, L., Gong, J., Yang, M., and Li, H. Omni-video: Democratizing unified video understanding and generation. *arXiv preprint arXiv:2507.06119*, 2025.

Team, D. Lucy edit: Open-weight text-guided video editing, 2025. URL https://d2drjpuinn46lb.cloudfront.net/Lucy_Edit__High_Fidelity_Text_Guided_Video_Editing.pdf.

Team, K., Chen, J., Ci, Y., Du, X., Feng, Z., Gai, K., Guo, S., Han, F., He, J., He, K., et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wang, Y., Yang, S., Zhao, B., Zhang, L., Liu, Q., Zhou, Y., and Xie, C. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033*, 2025.

Wei, C., Liu, Q., Ye, Z., Wang, Q., Wang, X., Wan, P., Gai, K., and Chen, W. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025.

Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., et al. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. 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*, pp. 7623–7633, 2023.

Wu, Y., Chen, L., Li, R., Wang, S., Xie, C., and Zhang, L. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16692–16701, 2025b.

Zhang, Z., Long, F., Li, W., Qiu, Z., Liu, W., Yao, T., and Mei, T. Region-constraint in-context generation for instructional video editing. *arXiv preprint arXiv:2512.17650*, 2025.

Zi, B., Ruan, P., Chen, M., Qi, X., Hao, S., Zhao, S., Huang, Y., Liang, B., Xiao, R., and Wong, K.-F. Se\~norita-2m: A high-quality instruction-based dataset for general video editing by video specialists. *arXiv preprint arXiv:2502.06734*, 2025.

A. Outlines

The supplementary material presents the following sections to strengthen the main manuscript:

- Section B details the Data Construction Process with Workflows.
- Section C details the Benchmark results.
- Section D provides additional Visualizations of the Reference dataset.
- Section E shows further Quantitative and Qualitative Comparisons with SoTA methods.

B. Dataset Details

The following prompt is used for MLLM grounding and Reference image socre filtering.

Prompt. 1: Editing Area Grounding

Given the image edit prompt and the source image and the edited image, generate the bounding box of the edited area in the edited image. Output format: { "image_2": [{"bbox": [x1, y1, x2, y2], "label": "object1"},...], }
Edit prompt is:

Prompt. 2: Foreground Grounding

Generate the bounding box of the same foreground object or person between the first and second images. Output format: { "image_1": [{"bbox": [x1, y1, x2, y2], "label": "object1"},...], "image_2": [{"bbox": [x1, y1, x2, y2], "label": "object1"},...], }

Prompt. 3: VLM Reference Image Score

You are an expert image editing evaluator. Given:
- an input edit instruction (add or replace),
- an input image,
- an edited image,
- a reference image representing the edited region,
evaluate whether the reference image is qualified.
Evaluation Criteria
1. Instruction-Follow Consistency
- The reference image must accurately represent the result of the input edit instruction as shown in the edited image.
- Object identity, attributes (color, shape, material, style), and edit type must be consistent.
- No contradictions with the edited result.
2. Image Quality & Focus
- Clear, realistic, and artifact-free.
- Coherent structure, plausible lighting and texture.
- The main subject must be clear and not overwhelmed by distracting elements.
Scoring
Output one overall score (1-10).
Final Output (JSON Only)
{"score": 1-10 integer}

C. Benchmark Details

The following prompt is used for RefVIE-Bench evaluation using Gemini (Comanici et al., 2025).

Prompt. 4: Evaluation Prompt of Subject Reference Editing Tasks

You are a data rater specializing in reference-guided object manipulation in videos. You will be given a Reference Image (the object to insert/swap), an Original Video, and the Edited Video. Your task is to evaluate the editing effect on a 5-point scale from three perspectives, specifically checking if the new object in the video matches the identity of the reference image.

Identity Consistency & Compliance

1. Object not swapped/added, or a completely unrelated object appears.
2. Object is changed, but looks nothing like the reference image (wrong color, shape, or class).
3. Object class is correct, but identity details (texture, specific markings, logos) differ significantly from the reference image.
4. High resemblance to the reference image; correct geometry and texture, with only minor variations in fine details.
5. Perfect identity transfer: The object in the video is indistinguishable from the reference image in terms of texture, structure, and style, while maintaining the correct pose for the scene.

Temporal Consistency & Texture Fidelity

1. The new object deforms, melts, or changes shape uncontrollably across frames.
2. Texture ”swims” or flickers; resolution drops significantly compared to the rest of the video; object vanishes in some frames.
3. Object is stable in form, but texture details blur or shift slightly during motion; style looks somewhat pasted-on.
4. Object is structurally solid and texture is consistent; minor edge shimmer or noise visible only on close inspection.
5. Completely temporally coherent; the object maintains rigid structure (or appropriate flexibility) and consistent texture details in every single frame, exactly like a real object.

Physical Integration & Tracking

1. Object slides around (bad motion tracking); does not follow camera or scene movement; looks like a sticker on the screen.
2. Missing interactions: No shadows, reflections, or occlusion handling (e.g., object appears on top of things that should be in front of it).
3. Motion tracking is decent with slight drift; lighting is flat or generic; occlusion is roughly correct but imprecise.
4. Accurate tracking; lighting and shadows match the scene’s direction and intensity; correct occlusion handling.
5. Physically flawless: Motion tracking, perspective changes, motion blur, shadows, reflections, and lighting interactions are indistinguishable from reality; the object feels physically present in the scene.

The second and third score should no higher than first score!!!

Example Response Format:

Brief reasoning: A short explanation of the score based on the criteria above, no more than 20 words.

Identity Consistency & Compliance: A number from 1 to 5.

Temporal Consistency & Texture Fidelity: A number from 1 to 5.

Physical Integration & Tracking: A number from 1 to 5.

editing instruction is : prompt Below are the reference image, original video, and edited video:

Prompt. 5: Evaluation Prompt of Background Reference Editing Tasks

You are a data rater specializing in video background replacement grading. You will be given a Reference Image, an Original Video (foreground subject), and the Edited Video (result). Your task is to evaluate the background replacement effect on a 5-point scale from three perspectives, paying close attention to the preservation of the foreground subject and the fidelity to the reference image.

Reference Fidelity & Preservation

1. Background not changed, or the foreground subject is severely damaged/removed.
2. Background changed but bears no resemblance to the reference image; foreground edges are significantly cut off or distorted.
3. Background resembles the reference but lacks key details; foreground is mostly preserved but has noticeable missing parts or artifacts.
4. Background clearly matches the reference image structure and style; foreground subject is fully preserved with only minor edge errors.

5. Perfect execution: The background is an exact semantic and stylistic match to the reference image, and the foreground subject is preserved pixel-perfectly throughout the entire duration.
Matting Quality & Temporal Stability
1. Severe flickering; the background or foreground jitters erratically; distinct ”boiling” artifacts on edges.
2. Obvious seams, halos, or ”green screen” outlines around the subject; background moves unnaturally or freezes while the camera moves.
3. Edges are generally stable but soft/fuzzy; minor flickering in complex areas (e.g., hair, transparent objects); background stability is acceptable.
4. Clean edges with minimal temporal noise; background motion aligns well with camera movement; casual viewers notice no matting errors.
5. Completely seamless composition; hair/transparency details are perfectly matted; background and foreground interact with perfect temporal stability in every frame.
Visual Harmony & Perspective
1. Background looks like a flat 2D image pasted behind a 3D subject; severe perspective or lighting mismatch (e.g., shadows point wrong way).
2. Lighting clashes (e.g., sunny background, dark foreground); no depth integration; subject looks ”floating.”
3. Perspective and scale are roughly correct; lighting is neutral but doesn't explicitly match the new environment's ambience.
4. Good environmental integration; foreground lighting tones reflect the new background; cast shadows are present and mostly accurate.
5. Photorealistic integration: Depth of field, motion blur, lighting, and color grading of the foreground perfectly match the reference background; the composite looks like a single, raw video capture.
The second and third score should no higher than first score!!!
Example Response Format:
Brief reasoning: A short explanation of the score based on the criteria above, no more than 20 words.
Reference Fidelity & Preservation: A number from 1 to 5.
Matting Quality & Temporal Stability: A number from 1 to 5.
Visual Harmony & Perspective: A number from 1 to 5.
editing instruction is : prompt Below are the reference image, original video, and edited video:

D. Sample Visualization

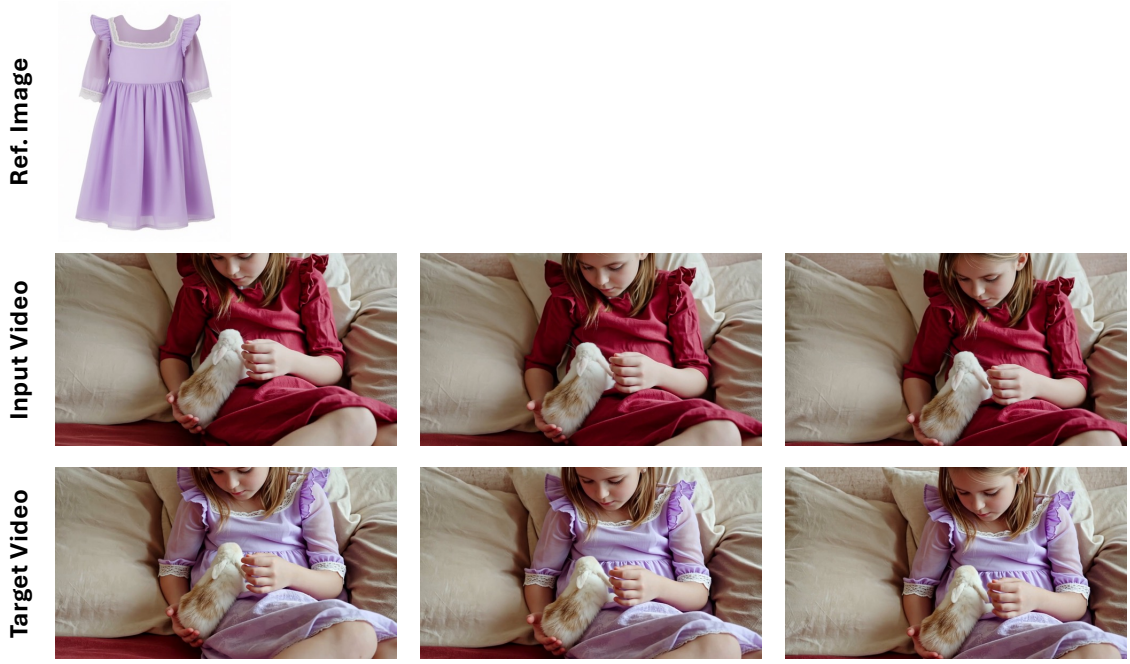
In this section, we present additional samples from our high-quality RefVIE, as illustrated in Fig. 8, Fig. 9, and Fig. 10.

E. Qualitative Comparison

This section compares our method with SoTA methods using instruction-only input (Figs. 11–12). *Instruction-reference-guided demo videos are provided in the supplementary material.*

Reference-guided Local Change

Instruction: *Replace the girl's red dress with a soft lavender one with lace trim and ruffled sleeves.*



Reference-guided Background Change

Instruction: *Replace the dynamic space station bay with the original static indoor car repair setting, featuring the open red car hood and consistent, even lighting.*

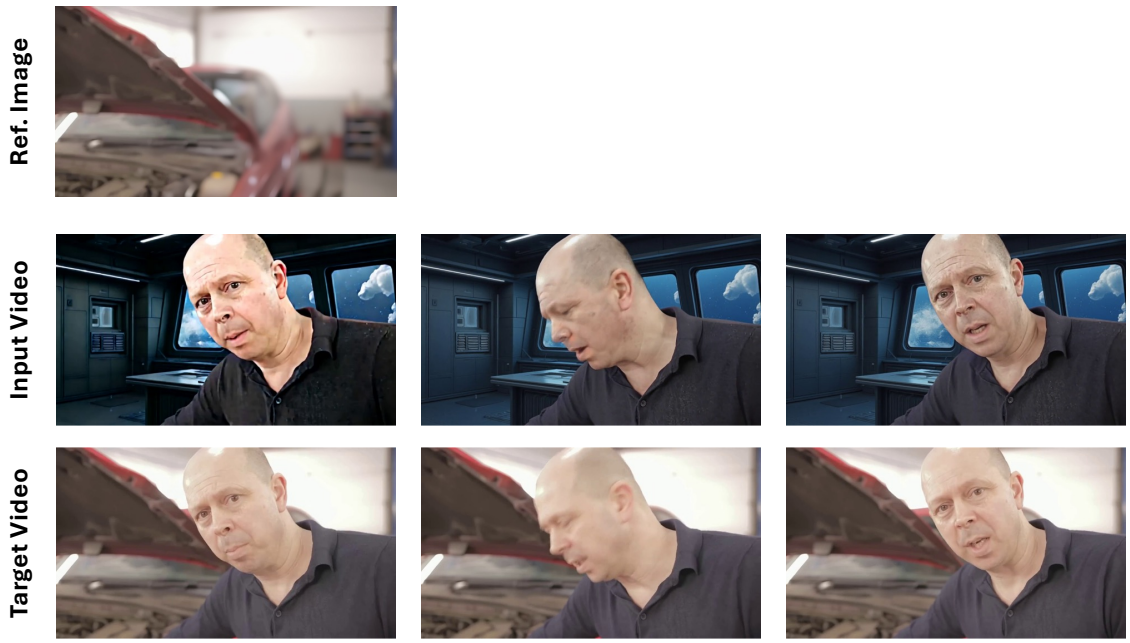
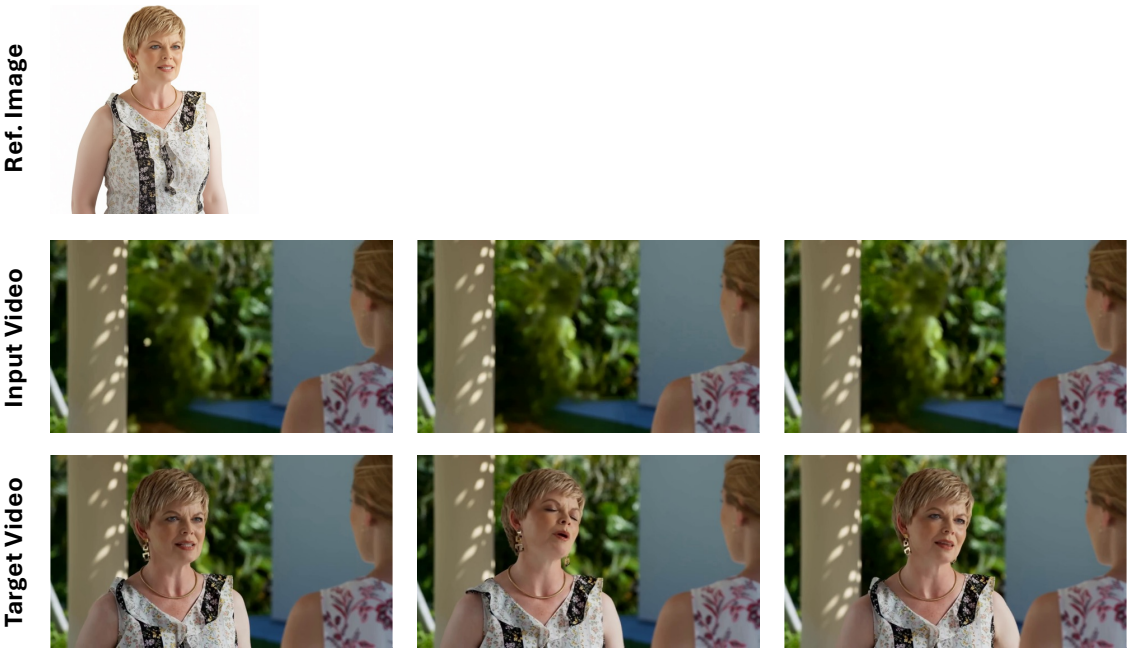


Figure 8. Examples of our RefVIE.

Reference-guided Local Add

Instruction: *Add the woman with styled blonde hair, sleeveless floral dress, delicate gold necklace, bright attentive eyes, subtle lip movements, and fluid natural movements into the video sequence. She must be tracked and integrated realistically and consistently across all frames.*



Reference-guided Change

Instruction: *Change the green valley to a golden wheat field under the sun.*

Figure 9. Examples of our RefVIE.

Reference-guided Local Change

Instruction: *Replace the dark intricately carved wooden structure on the left with a woman in a purple and blue outfit with bangs, looking right.*

Reference-guided Local Add

Instruction: *Add a deep purple scarf draping over the shoulders into the video, ensuring the scarf is tracked and animated consistently and realistically throughout the sequence.*

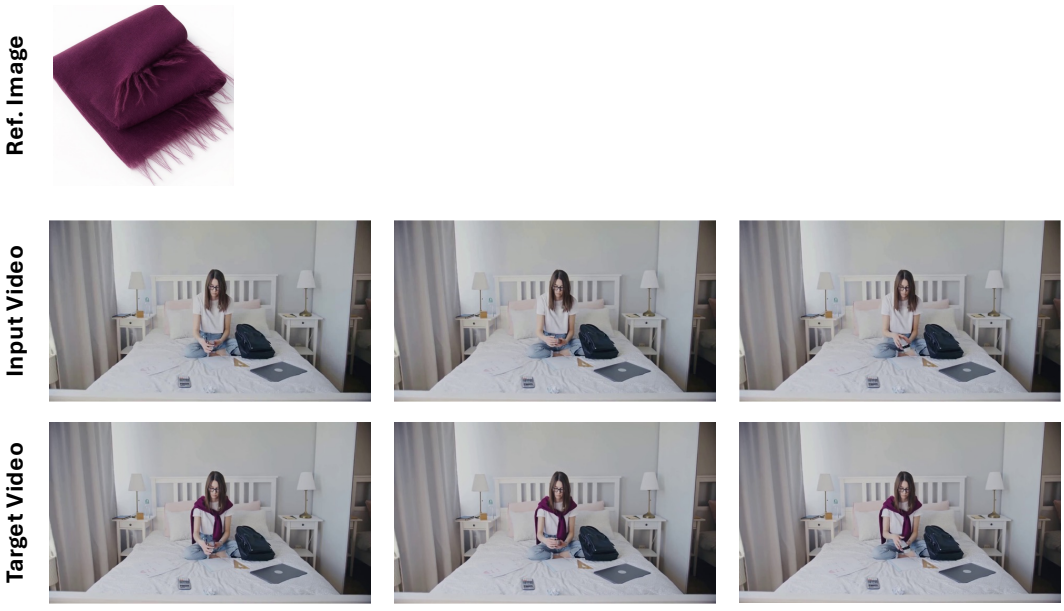


Figure 10. Examples of our RefVIE.

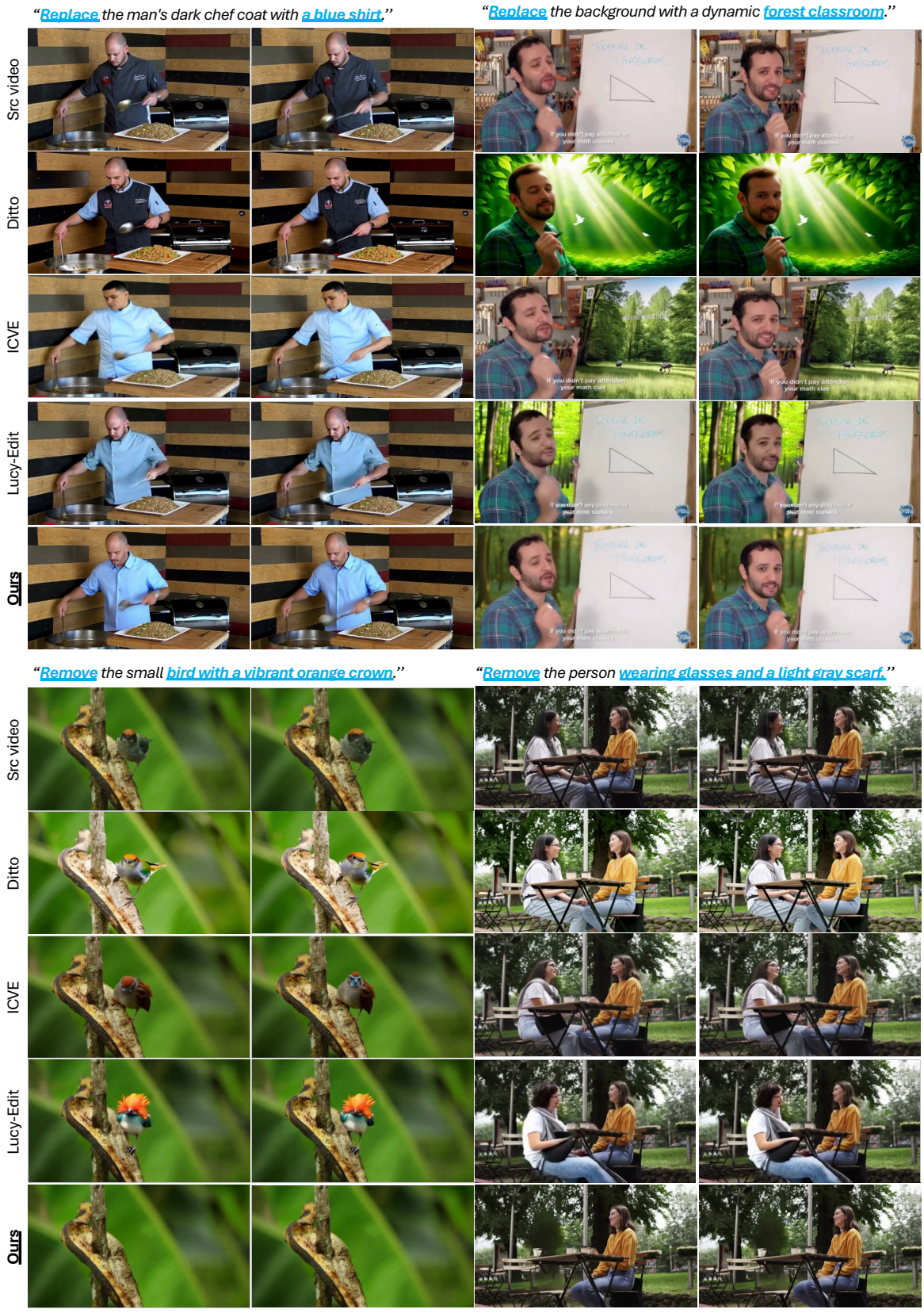


Figure 11. Visual comparison on OpenVE-Bench.

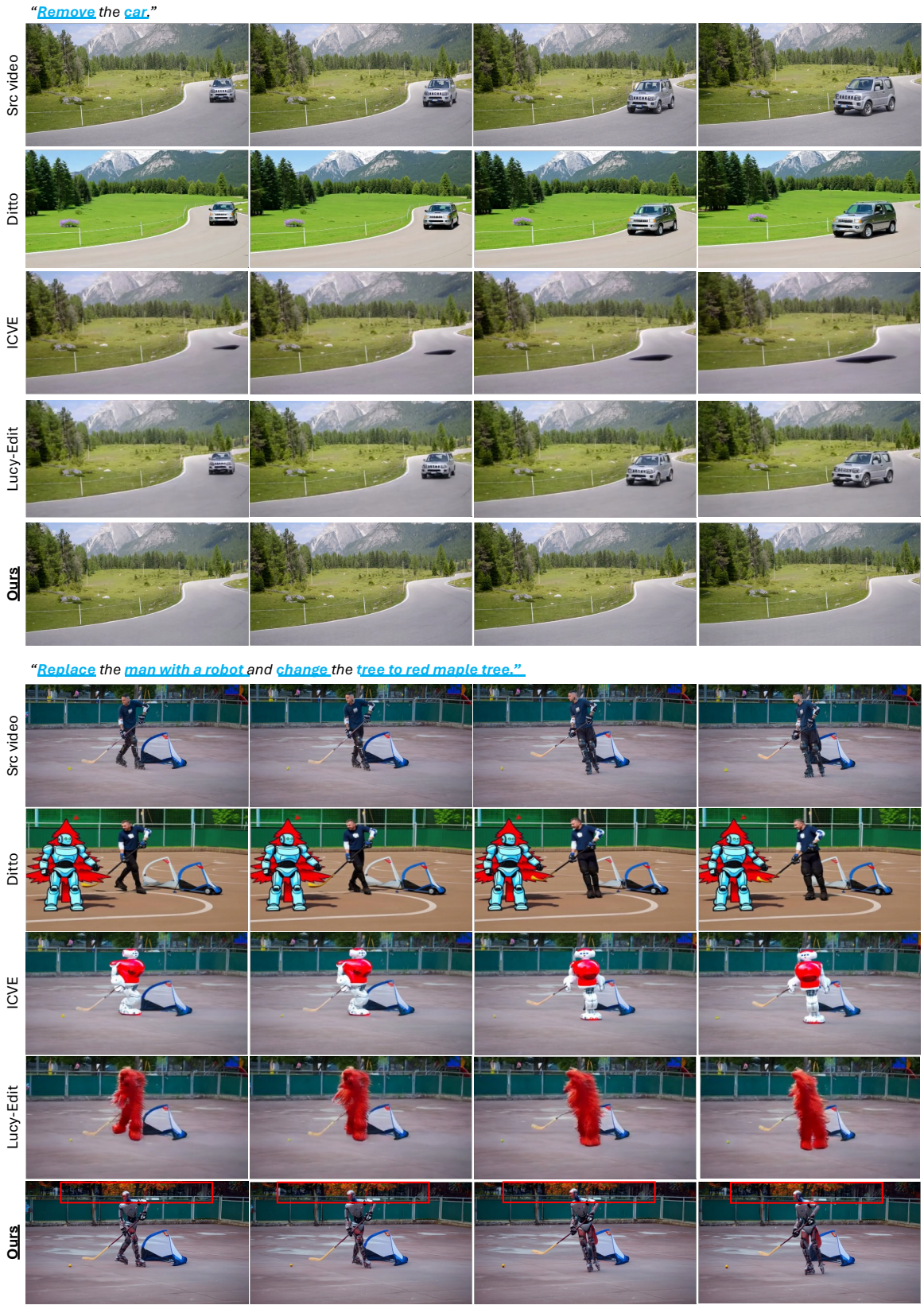


Figure 12. Visual comparison on VIE-Bench. The bottom instruction not only replaces the man with a robot but also changes the tree in the background to a red maple tree. Only our method precisely follows this instruction, as highlighted in the red box.