FIRST-FRAME-GUIDED VIDEO EDITING WITH A VISUAL AUTOREGRESSIVE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

While *text-to-video* generative models have advanced video editing, maintaining spatiotemporal consistency in direct video manipulation remains challenging. As a practical alternative to direct video editing, first-frame guided video editing has emerged as a new paradigm, which leverages the superior fidelity of image editing by propagating edited attributes from the first frame to the full sequence. Existing first-frame guided video editing methods are primarily built on diffusion and flow matching models, which struggle to disentangle appearance from motion. To tackle this problem, we propose a novel framework that explicitly exploits the inherent inductive bias of *visual autoregressive* (VAR) models. Unlike diffusion and flow-matching models that adapt appearance and motion on a shared fixed-resolution latent space, VAR exposes an explicit coarse-to-fine hierarchy, making motion easier to disentangle from the appearance during adaptation. Exploiting this structure, we restrict motion adaptation to low scales while preserving high-scale appearance refinement, achieving high-fidelity first-frame guided video editing with robust source motion preservation. Extensive evaluations demonstrate that our approach achieves leading input-frame consistency while maintaining strong source-motion preservation across both local and global edits.

1 INTRODUCTION

Recent advances in generative models (Ho et al., 2020; Song et al., 2021; Peebles & Xie, 2023; Rombach et al., 2022; Hong et al., 2023; Yang et al., 2025; Zhang et al., 2024; Kong et al., 2024; Wan et al., 2025) have driven significant progress in high-quality video generation. Building on this, video editing (Wu et al., 2023a; Song et al., 2024; Geyer et al., 2024; Jiang et al., 2025) has also continuously improved. However, despite these improvements, directly editing video often yields suboptimal editing results and limited controllability, even at substantial computational cost. In contrast, single-image editing models (Brooks et al., 2023; Hertz et al., 2023; Mao et al., 2025; Kulikov et al., 2025) have already achieved exceptional performance and fine-grained controllability. To leverage these advantages, first-frame guided video editing (Ouyang et al., 2024; Burgert et al., 2025; Gao et al., 2026) has emerged as a highly practical alternative. Specifically, as illustrated in Fig. 1, given an edited initial frame shown in the first column, the task aims to propagate its modified appearance and style to the remaining frames of a source video. By leveraging high fidelity of single-image editing, this approach successfully circumvents the difficulties of direct video manipulation.

In first-frame guided video editing, the key objective is to naturally propagate the edited first frame across the original video. It remains a critical challenge to apply the newly modified style and content while faithfully preserving the unique temporal dynamics of the source video. Most recent methodologies (Ouyang et al., 2024; Burgert et al., 2025; Gao et al., 2026) have relied on diffusion models to tackle this problem. For instance, I2VEdit (Ouyang et al., 2024) attempts to propagate first-frame edits using Image-to-Video (I2V) diffusion models, utilizing Low-Rank Adaptation (LoRA) (Hu et al., 2022) and skip-interval cross attention to maintain motion fidelity. However, since the appearance features are deeply entangled with the motion, the edited results become degraded, leading to severe visual artifacts and color saturation. This stems from the inherent property of diffusion models, which operate at a shared fixed-resolution throughout the entire denoising process. Throughout this iterative generation, motion and appearance features remain deeply entangled within the same latent space, significantly hindering independent control. To circumvent this inherent entanglement, LoRAEdit (Gao et al., 2026) relies on spatial masks to preserve the original motion,

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.

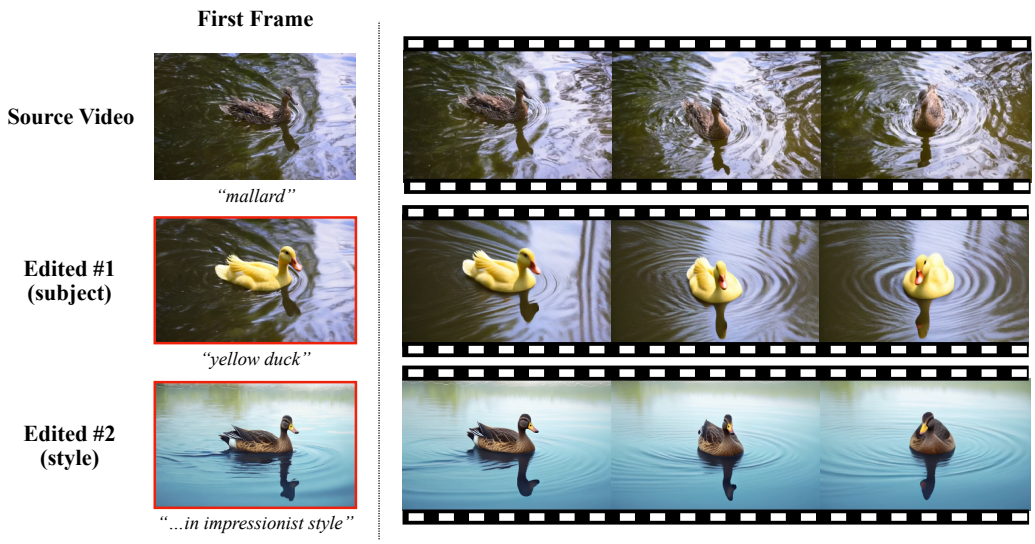

Figure 1: **First-Frame Guided Video Editing.** Given a source video (top row) and its corresponding first frame (leftmost column), the task is to propagate the modified visual attributes across the entire temporal sequence. The second and third rows illustrate the propagation of localized and global edits given the edited first frames (leftmost column), respectively.

yet such mask-dependency often introduces temporal flickering along the boundaries and struggles to accommodate significant structural changes or global edits involving the entire background.

To overcome the structural limitations of diffusion-based approaches, we propose a novel video editing framework leveraging the newly emerged next-scale prediction paradigm of *visual autoregressive* (VAR) models (Liu et al., 2025a). Rather than relying on heuristic masks or complex parameter scaling, we exploit VAR's explicit generation-scale hierarchy, which makes temporal dynamics easier to disentangle from appearance than in diffusion-based approaches. Specifically, InfinityStar (Liu et al., 2025a), which is a *text-to-video* VAR model, explicitly separates a motion-free T2I regime from a motion-generating I2V process, allowing appearance to be learned independently of temporal dynamics. Within the I2V stage, the coarse temporal structure is predominantly captured at early low-resolution scales, while fine-grained visual details are progressively synthesized in the subsequent high-resolution scales. Although video diffusion models also exhibit an implicit motion-to-detail progression across denoising timesteps (Wu et al., 2025; Baherwani et al., 2025), their denoising state remains on a fixed-resolution latent grid. VAR instead realizes this hierarchy through explicit token maps at distinct resolutions, making appearance–motion disentanglement more direct and enabling source-motion adaptation only at low-resolution I2V scales.

Inspired by two-stage LoRA (Hu et al., 2022) techniques in diffusion-based video editing models (Ren et al., 2024; Abdal et al., 2025; Ma et al., 2026), we reformulate this approach for the multi-scale architecture of VAR (Tian et al., 2024; Liu et al., 2025a), introducing a scale-specific strategy to decouple appearance and motion. In Stage 1, we train an Appearance LoRA in the motion-free T2I regime using a single reference image (the first frame of the source video) associated with an initial abstract text token to learn the appearance. In Stage 2, we freeze the Appearance LoRA and train a Motion LoRA on the lower resolution video scales guided by a distinct placeholder text token to learn source dynamics. As an auxiliary refinement, we incorporate an optical-flow-based regularization to guide the model toward better temporal alignment, primarily in sequences with complex motion, without limiting the flexibility of the appearance transfer. Our GT-prefix analysis and scale-gated two-stage adaptation establish a VAR-specific framework for first-frame-guided video editing, achieving state-of-the-art input consistency and strong source-motion preservation across local and global edits.

2 Related Work

Visual Autoregressive Model. While diffusion models (Ho et al., 2020; Peebles & Xie, 2023; Song et al., 2021) have driven significant progress in Text-to-Image (T2I) (Chen et al., 2024; 2023; Kong et al., 2024; Zhang et al., 2024; Nichol et al., 2022; Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Podell et al., 2024; Li et al., 2024; Xie et al., 2025) and Text-to-Video (T2V) (Hong et al., 2023; Yang et al., 2025; Zhang et al., 2024; Kong et al., 2024; Wan et al., 2025) generation by achieving exceptional visual fidelity, their iterative denoising process incurs substantial latency. To resolve these limitations, recent VAR models (Tian et al., 2024; Han et al., 2024; Liu et al., 2025a) introduce a next-scale prediction paradigm. By progressively generating features from coarse to fine resolutions, VAR enables efficient parallel decoding while preserving structural integrity. Building upon this paradigm, recent works like Infinity (Han et al., 2024) and InfinityStar (Liu et al., 2025a) have successfully extended next-scale prediction to foundational T2I and T2V generation. Crucially, this hierarchy explicitly separates global-structure and fine-detail prediction, providing a favorable scale axis for disentangling temporal dynamics from spatial appearance.

First-Frame guided Video Editing. While video editing (Wu et al., 2023a; Song et al., 2024; Geyer et al., 2024) often struggles to maintain strict spatiotemporal consistency, first-frame guided video editing (Ouyang et al., 2024; Liu et al., 2025b; Gao et al., 2026) has emerged as an elegant alternative. By propagating high-fidelity edits from a single initial frame across the sequence, it effectively bypasses temporal artifacts. In parallel, several methods (Huang et al., 2025; Jiang et al., 2025) adopt reference-guided video editing frameworks, which conditionally incorporates a modified image to guide the video editing process. However, because these reference-guided models are not explicitly optimized for continuous single-frame propagation, they often suffer from temporal degradation. Furthermore, some of these methods (Jiang et al., 2025; Gao et al., 2026) rely on spatial masks to ensure temporal consistency, which inherently limits their scope to local editing.

3 Method

3.1 Preliminaries: Visual Autoregressive Video Generation

InfinityStar. InfinityStar (Liu et al., 2025a) is a unified spacetime autoregressive Text-to-Video model by sequentially performing Text-to-Image (T2I) and Image-to-Video (I2V) synthesis using next-scale prediction. The underlying architecture models the transition from text to a dynamic sequence through a hierarchical Spacetime Pyramid structure. Specifically, the continuous visual feature F_k^c at scale k within the c-th clip (where clip denotes the temporal segment) is progressively constructed by accumulating upsampled multi-scale residual maps R_i^c. To predict the subsequent residual token R_k^c, the accumulated feature F_{k-1}^c is then downsampled to provide spatial conditioning $\tilde{F}_{k-1}^c$ to match the required input shape for the model:

$$F_k^c = \sum_{i=0}^{k} \text{up}(R_i^c), \qquad \text{and} \qquad \tilde{F}_{k-1}^c = \text{down}(F_{k-1}^c) \tag{1}$$

where $\text{up}(\cdot)$ and $\text{down}(\cdot)$ represent the upsampling and downsampling operations, respectively.

While this scheme operates within the 3D spacetime pyramid of a single clip, inter-clip information is propagated by passing the finest resolution features of the preceding clip $(c-1)$ as the Key and Value states for the attention layers of the current clip (c). Specifically, since the generation process begins with the T2I stage as the 0-th clip ($c = 0$), followed by the I2V stage as the 1-st clip ($c = 1$), the generated image serves as a reference for the subsequent video generation clip. For a video sequence comprising N clips and K resolution scales, the autoregressive likelihood is formulated as:

$$p(R_0^0, \ldots, R_{K-1}^{N-1} \mid \psi(t)) = \prod_{c=0}^{N-1} \prod_{k=0}^{K-1} p(R_k^c \mid R_{<k}^c, R_{0:K-1}^{<c}, \psi(t)). \tag{2}$$

where R_k^c denotes the k-th residual token of the c-th clip, and $\psi(t)$ represents the text conditioning. To efficiently model this likelihood, InfinityStar (Liu et al., 2025a) employs Binary Spherical Quantization (BSQ) (Zhao et al., 2025), which projects the latent representations into a binary space. Consequently, predicting the residual token R_k^c is reduced to channel-wise binary classifications.

Similar to DiT-based Text-to-Video models (Wan et al., 2025; Yang et al., 2025), InfinityStar (Liu et al., 2025a) utilizes a full 3D attention within its transformer blocks, enabling the joint capturing of both spatial and temporal dependencies. Due to a significant disparity in visual fidelity among available weights, we adopt the InfinityStar-720p-8B model as it is the only version capable of producing reliable results. As this specific model is constrained to $c = 1$ (capable of generating up to 81 frames), we restrict our study to $c \in \{0, 1\}$, excluding long video editing.

Ground Truth Token Injection. InfinityStar (Liu et al., 2025a) leverages the property that ground-truth (GT) tokens can be easily obtained by encoding any given video or image through its visual tokenizer. It allows for the direct injection of GT tokens into the residual maps at each scale. By utilizing it, if a specific image or video is provided, the model can extract the corresponding GT tokens and assign the residual maps up to a desired scale. The subsequent scales are then generated based on these fixed priors. In our framework, it is particularly effective when an edited image is available. The GT tokens of the edited image are extracted and used as residual maps to generate a video sequence conditioned on that image as its starting point. By passing these injected GT tokens through the transformer, the model can perform KV caching, ensuring that the inference process operates identically to its standard sampling process while maintaining high fidelity to the reference input.

3.2 Analysis: Motion Reconstruction via Low-scale Features

Given VAR's explicit coarse-to-fine generation hierarchy, we hypothesize that source motion is concentrated at lower-resolution scales, whereas higher scales primarily refine spatial details. Using frozen base InfinityStar (Liu et al., 2025a), we test this hypothesis by measuring how much low-scale ground-truth (GT) information is required to reconstruct source motion. Specifically, we conduct an experiment in which visual tokens at each scale prior to scale k are fixed as GT, while the remaining scales are predicted by the model. By varying the exclusive cutoff k_c from 0 to $K - 1$, we examine the resulting videos to observe how the degree of GT guidance affects motion fidelity. To quantitatively evaluate motion preservation, we adopt the Motion Fidelity metric introduced by Yatim et al. (2024). Given CoTracker3 (Karaev et al., 2025) trajectories from the source and generated videos, we normalize their frame-to-frame displacement vectors and compute the temporal cosine similarity between each pair of trajectories.

Table 1: **Motion Reconstruction Analysis.** S_{k_c} denotes the Motion Fidelity between the source and generated videos when GT tokens are injected at scales $s < k_c$ (i.e. the first k scales). The scores are averaged across all videos used in our main evaluation.

k_c	$\lvert S_{k_c}/S_{recon}\rvert$	k_c	$\lvert S_{k_c}/S_{recon}\rvert$	k_c	$\lvert S_{k_c}/S_{recon}\rvert$
0	0.762	5	0.938	10	0.975
1	0.782	6	**0.951**	11	0.981
2	0.842	7	0.960	12	0.990
3	0.897	8	0.964	13	0.994
4	0.928	9	0.969	14	0.997

To account for VAE reconstruction loss, Tab. 1 reports S_k/S_{recon}, where S_{recon} is measured from the reconstruction obtained by injecting GT tokens at all scales. At $k = 6$, S_k/S_{recon} reaches 0.951, indicating that the low-resolution scales retain over 95% of the recoverable source motion. This early concentration of motion along VAR's explicit generation scales provides a clearer separation axis than the fixed-resolution latent spaces of diffusion and flow-matching models, making appearance and motion easier to disentangle. We therefore use an exclusive training cutoff $k_c = 6$, adapting the first six scales.

3.3 Two-Stage Training Strategy

While Sec. 3.2 confirms that lower scales primarily encapsulate temporal dynamics, these scales are not entirely 'motion-pure'; they still retain coarse spatial representations, although VAR's explicit organization across coarse-to-fine scales makes the underlying appearance and motion significantly easier to disentangle than the densely coupled features in diffusion and flow-matching models. If trained jointly, the model often struggles to distinguish between the two, leading to appearance leakage where the motion adapters inadvertently overfit to static textures, ultimately causing temporal flickering or shape distortion. Leveraging the structural advantage discussed in Sec. 3.1, where InfinityStar (Liu et al., 2025a) performs motion-free T2I generation followed by I2V generation,

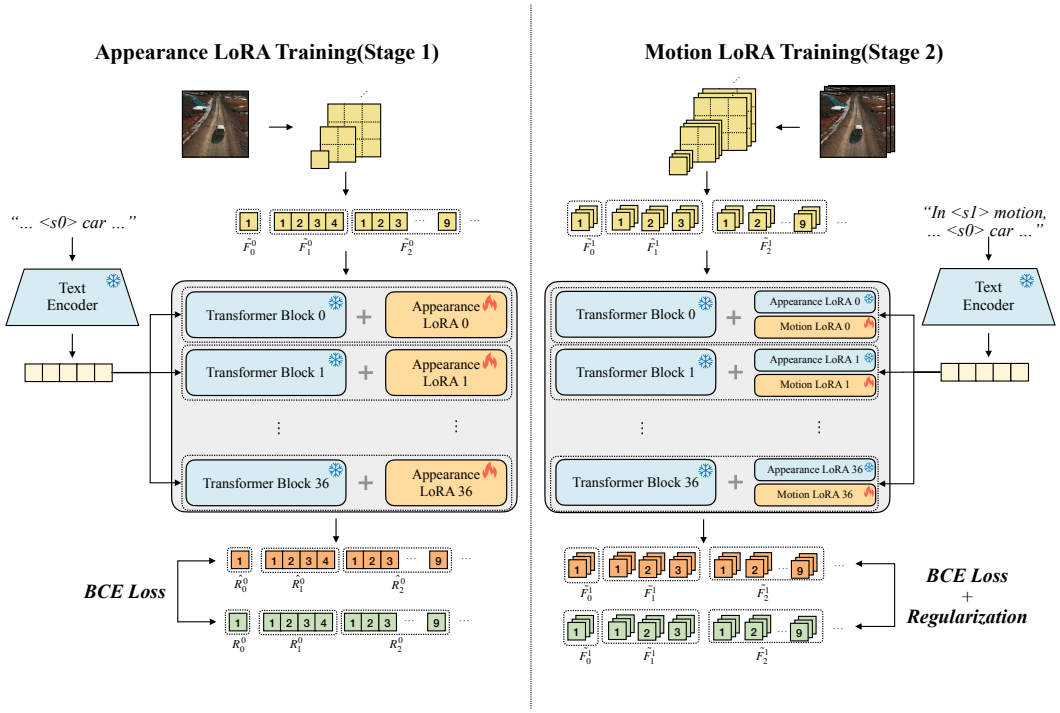

Figure 2: **Method Overview.** (Left) In Stage 1, the Appearance LoRA (Hu et al., 2022) is trained using only the first frame encoded by the visual tokenizer, guided by the BCE Loss ($\mathcal{L}_{BCE}$). (Right) In Stage 2, the Appearance LoRA is frozen, and the Motion LoRA is trained on the full video sequence using a combination of the BCE Loss ($\mathcal{L}_{BCE}$) and the Flow-Guided Regularization ($\mathcal{L}_{motion}$).

we reformulate prior diffusion-based two-stage adaptation strategies for the VAR framework by exploiting VAR's multi-scale hierarchy structure.

In Stage 1, operating in the motion-free T2I phase, we explicitly capture and 'lock' the spatial priors within an Appearance LoRA (Hu et al., 2022). By subsequently freezing these parameters, we create a structural bottleneck that forces the Motion LoRA in Stage 2, trained in the I2V phase, to focus exclusively on distilling temporal dynamics, effectively preventing any interference from the underlying appearance signals. The overview of our two-stage training framework is illustrated in Fig. 2.

3.3.1 Stage 1: Appearance LoRA Training

In the first stage, we focus on capturing the visual appearance of the source video by adapting the model to its representative frame. Leveraging the structure of successive T2I and I2V pipeline in InfinityStar (Liu et al., 2025a), we find that training the Appearance LoRA (Hu et al., 2022) solely on the T2I stage using the first 6 scales of a single image is sufficient to acquire the necessary spatial features. This approach offers significant computational efficiency, as processing tokens for a static image instead of a full video sequence drastically reduces the total token count. Specifically, we integrate LoRA modules into the query (q), key (k), value (v), and output (o) projections of the attention blocks, as well as the gate and down projections of the MLP blocks. To guide the model in encoding the visual appearance into the generation process, we prepend a meaningless special token (e.g. "[s0]") before the subject in the text prompt. Crucially, while the embedding of the special token remains frozen, the Appearance LoRA is exclusively trained to encapsulate the appearance in video. The optimization is guided by the scale-wise Binary Cross-Entropy loss (BCE Loss):

$$\mathcal{L}_{\text{Stage1}} = -\sum_{k=0}^{5} \left[R_k^0 \log \hat{R}_k^0 + (1 - R_k^0) \log(1 - \hat{R}_k^0) \right]. \tag{3}$$

where R_k^0 represents the GT binary token at scale k within the image-level scale maps in T2I clip ($c = 0$), and $\hat{R}_k^0$ is the predicted probability conditioned on the preceding scales $R_{<k}$ and the text condition $\psi(t)$ containing the special token. Note that the optimization in Stage 1 is restricted to the image scales of the T2I phase.

3.3.2 STAGE 2: MOTION LORA TRAINING

While Stage 1 is strictly designed to learn the appearance information and establish the spatial priors, Stage 2 focuses on learning temporal dynamics by training a separate Motion LoRA (Hu et al., 2022) that is effectively disentangled from the established spatial representations. As analyzed in Sec. 3.2, the first six resolution scales ($k = 0, ..., 5$) sufficiently capture the core temporal dynamics. Therefore, we restrict the motion training to these specific scales, effectively maintaining motion fidelity while significantly reducing computational overhead.

The objective for Stage 2 combines the standard BCE Loss, extended from Stage 1 across frames, with auxiliary flow-guided regularization term. For the I2V stage, $\hat{R}_k^1$ additionally conditions on the T2I tokens $R_{0:K-1}^0$, yielding:

$$\mathcal{L}_{\text{BCE}} = -\sum_{k=0}^{5} \left[R_k^1 \log \hat{R}_k^1 + (1 - R_k^1) \log(1 - \hat{R}_k^1) \right]. \tag{4}$$

Importantly, the accumulated spatial visual features at the finest resolution–which inherently encapsulate the residual information across all scales in T2I stage($R_{1:K}^0$)– are directly injected as Key and Value states within the attention layers.

Auxiliary Flow-Guided Regularization. While $\mathcal{L}_{BCE}$ exhibits a certain level of generative capability in synthesizing temporal sequences, we empirically observe that relying solely on this objective yields sub-optimal motion dynamics in some challenging cases, particularly for localized or complex human motion. We suspect that the BCE Loss, by uniformly penalizing token-level discrepancies across all locations, can underemphasize sparse motion-bearing regions relative to static spatial content. To address it, we introduce a lightweight flow-guided regularization term ($\mathcal{L}_{motion}$) utilizing optical flow extracted from the source video to improve robustness to such challenging dynamics during the training. For each scale $k \in \mathcal{K}_m = \{2, 3, 4, 5\}$, backward optical flow is estimated using RAFT (Teed & Deng, 2020). Every four consecutive flows are spatially composed and downsampled to align with the resolution and compressed frame rate. We form a detached binary proxy $\mathbf{R}_{k,t}^1$ from the current-frame scale-matched head logits $\widehat{\mathbf{R}}_{k,t}^{1,\text{logit}}$ and warp it to supervise the raw next-frame logits over moving regions:

$$\mathcal{L}_{\text{motion}} = \frac{1}{|\mathcal{K}_m|(T-1)} \sum_{k \in \mathcal{K}_m} \sum_{t=1}^{T-1} \text{BCE}\left(\widehat{\mathbf{R}}_{k,t+1}^{1,\text{logit}}, \text{Warp}\left(\mathbf{R}_{k,t}^1, \mathcal{V}_{k,t+1 \to t} \right) \right). \tag{5}$$

where T is the number of temporally compressed frames, BCE denotes binary cross-entropy with logits evaluated over moving regions, and $\widehat{\mathbf{R}}_{k,t}^{1,\text{logit}}$ denotes the pre-sigmoid logits predicted by the scale-matched semantic or detail head at frame t. Here, $\mathcal{V}_{k,t+1 \to t}$ denotes the spatially composed backward flow at scale k. Note that the coarsest scales ($k = 0, 1$) are excluded from this regularization due to their insufficient spatial resolution for meaningful warping. Consequently, the overall optimization objective for Stage 2 is formulated as:

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{BCE}} + \alpha \cdot \mathcal{L}_{\text{motion}} \tag{6}$$

where α is a balancing hyperparameter.

3.4 INFERENCE

Given an edited first frame, we encode it into GT visual tokens and directly inject them into the T2I scale maps. Since the target appearance is provided by these injected tokens, the Appearance LoRA (Hu et al., 2022) is disabled, and only the source-specific Motion LoRA is used during I2V generation. The Motion LoRA is activated at scales $k < 6$ and gated to zero at all subsequent scales, allowing the frozen backbone to refine high-resolution details without additional appearance interference.



Figure 3: **Qualitative Results.** The top rows illustrate local editing, and the bottom present global. For global editing, mask-based methods are excluded. Ours faithfully propagates edited attributes while preserving original motion, mitigating artifacts or motion inconsistency observed in others.

4 Experiments

4.1 Implementation Details

We implement our framework based on the InfinityStar-8B 720p model (Liu et al., 2025a). Due to the architectural constraints of the InfinityStar, all experiments are conducted using video sequences of 81 frames. Training is performed for 300 steps in Stage 1 and 400 steps in Stage 2, where both stages are learned over scales $k \in \{0, 1, \ldots, 5\}$. We set learning rate to 1×10^{-4} for both stages and the regularization weight α to 0.001.

Comparison. We compare our method against several state-of-the-art frameworks, including I2VEdit (Ouyang et al., 2024), Go-with-the-Flow (Burgert et al., 2025), VACE (Jiang et al., 2025), and LoRAEdit (Gao et al., 2026). While VACE offers various editing capabilities, in our task, we specifically utilize its mask-based, reference-guided video editing for the comparison. To adapt VACE

Table 2: Quantitative comparison on local and global editing. Best and second-best results within each setting are **bolded** and underlined, respectively.

Edit	Method	Visual Quality		Input Consistency			Temporal Consistency
		Imaging	Aesthetic	Subject	Background	CLIP-I2V	Motion Fidelity
Local	I2VEdit	0.6214	0.5549	0.9737	0.9651	0.8671	**0.8319**
	VACE	**0.6781**	0.5676	0.9561	0.9482	0.8721	0.7915
	Go-with-the-Flow	0.6608	**0.5718**	0.9523	0.9446	0.8750	0.7371
	LoRAEdit	0.6663	0.5616	0.9511	0.9409	0.8684	0.7868
	Ours	0.6621	0.5702	**0.9753**	**0.9656**	**0.8933**	0.8110
Global	I2VEdit	0.5741	0.5783	0.9684	0.9613	0.8652	0.7245
	Go-with-the-Flow	0.6283	0.5940	0.9474	0.9526	0.8784	0.7198
	Ours	**0.6503**	**0.6071**	**0.9701**	**0.9633**	**0.8831**	**0.7252**

to our experimental setup, we provide the first frame as the reference image and employ SAM 3 (Carion et al., 2026) to generate the masks for the target objects. LoRAEdit and VACE rely on explicit masks and are therefore evaluated only on local edits, as full-frame masks nullify their mask-based preservation mechanism.

Evaluation Metrics. To comprehensively evaluate our method, we categorize metrics into three distinct categories: visual quality, input consistency, and temporal consistency. We utilize VBench (Huang et al., 2024a) **Imaging Quality** and **Aesthetic Quality** to assess synthesis quality. To measure spatial alignment with the first frame, we employ VBench++ (Huang et al., 2024b) **Subject Consistency**, **Background Consistency** for I2V, as well as a CLIP-based (Radford et al., 2021) I2V score. Finally, we adopt the **Motion Fidelity** from (Yatim et al., 2024), leveraging CoTracker3 (Karaev et al., 2025) to measure how well the original motion is preserved.

Dataset. To conduct a quantitative evaluation, we utilized a total of 60 videos collected from I2VEdit (Ouyang et al., 2024), loveu-tgve-2023 dataset (Wu et al., 2023b), and IVEBench (Chen et al., 2026). Due to architectural requirements of InfinityStar (Liu et al., 2025a), we specifically selected videos with a minimum length of 80 frames. For samples without edited first frame, we prepared the reference first frames using ACE++ (Mao et al., 2025) for object-level manipulation, and ControlNet (Zhang et al., 2023) with Flux.1-dev (Labs, 2024) for stylistic and background transformations.

4.2 Results

We comprehensively evaluate ours against state-of-the-art baselines through qualitative results, quantitative results, and user study. To ensure a fair comparison, VACE (Jiang et al., 2025) and LoRAEdit (Gao et al., 2026) are included only in the local editing results, as they are inherently restricted to mask-guided editing.

Qualitative Results. We present the qualitative results in Fig. 3. I2VEdit (Ouyang et al., 2024) adheres to the target motion in the local editing examples, yet it eventually suffers from severe pixel collapse and saturation. It further fails to maintain motion fidelity in global editing. Go-with-the-Flow (Burgert et al., 2025) demonstrates limited capability in capturing underlying motion dynamics, leading to significant temporal inconsistency. Although VACE (Jiang et al., 2025) and LoRAEdit (Gao et al., 2026) utilize explicit masks, both methods fail to preserve consistency with the reference frame over time, as shown in the clear mismatch with the edited first frame, such as the color inconsistency in the bus or the disappearance of the hat. In contrast, our method faithfully follows the intended motion and maintains high consistency with the edited first frame in both settings, demonstrating superior structural and temporal integrity. Additional results are provided in the supplementary material.

Quantitative Results. As presented in Tab. 2, our method achieves the strongest overall performance, ranking first on all input consistency metrics and second on motion fidelity for local edits. I2VEdit (Ouyang et al., 2024) attains higher motion fidelity but substantially lower imaging quality. We attribute it to motion coupling in its fixed-resolution diffusion latent, where source dynamics and appearance remain coupled and induce saturation during edit propagation shown in Fig. 3. For global edits, ours obtains the highest score on all metrics. VACE (Jiang et al., 2025) and LoRAEdit (Gao et al., 2026) achieve higher local imaging scores which stems from the use of explicit masks to

Table 3: User Study results on Local and Global Editing. Lower scores indicate better performance.

Method	Local Editing		Global Editing	
	Input Consistency ↓	Video Fidelity ↓	Input Consistency ↓	Video Fidelity ↓
VACE	4.284	2.889	–	–
LoRAEdit	2.259	2.580	–	–
I2VEdit	3.580	4.235	2.404	2.543
Go-with-the-Flow	2.914	3.309	2.394	2.287
Ours	**1.963**	**1.988**	**1.202**	**1.170**

preserve unedited pixels, whereas ours achieves comparable score while supporting both local and global edits without masks.

User Study. We conduct a user study evaluating Input Consistency and Video Fidelity across both local and global editing scenarios. Participants rank the results of each method, and final scores are calculated as the average ranking, where a lower value implies higher quality. As shown in Tab. 3, our method consistently achieves the best average rankings, outperforming existing baselines.

Table 4: Ablation study on the effect of flow-guided regularization.

Method	Visual Quality		Input Consistency			Temporal Consistency
	Imaging	Aesthetic	Subject	Background	CLIP-I2V	Motion Fidelity
w/o Regularization	**0.6576**	**0.5906**	**0.9735**	0.9635	0.8867	0.7569
w/ Regularization	0.6560	0.5892	0.9726	**0.9644**	**0.8880**	**0.7667**

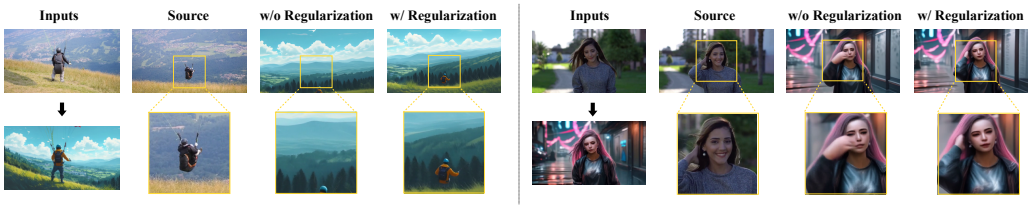

Figure 4: **Ablation Study on the auxiliary Flow-Guided Regularization.** The regularization helps preserve motion consistency in challenging scenes, particularly under global edits with human posing.

4.3 Ablation Study

To validate the effectiveness of the Flow-Guided regularization, we conduct an ablation study using the same dataset from our main evaluation. As shown in Tab. 4, while both models achieve comparable performance in visual quality and input consistency, incorporating regularization leads to an improvement in motion fidelity. Empirically, although the difference between the two configurations seems marginal, the model with regularization demonstrates superior temporal consistency in complex scenes particularly in global edits involving human motion as shown in Fig. 4.

5 Conclusion

In this paper, we present a novel framework for first-frame guided video editing that exploits two structural advantages of VAR: a motion-free T2I regime and an explicit coarse-to-fine I2V hierarchy. Together, these properties make appearance and motion substantially easier to disentangle than in fixed-resolution diffusion and flow-matching models. Accordingly, we propose a two-stage LoRA strategy. Stage 1 learns appearance on image-level scales, while Stage 2 learns temporal dynamics via a video-level motion LoRA, supplemented by auxiliary flow-guided regularization. Extensive experiments show that ours achieves the strongest mean input consistency and a favorable overall balance between motion fidelity and visual quality.

AI USE STATEMENT

Generative AI tools were used to assist with translation, language polishing, literature discovery, manuscript editing, result interpretation, code refinement and debugging, and to provide feedback on experimental design. AI-based image-editing tools are additionally used to generate edited first-frame inputs for evaluation as described in Sec.4.1. The core research idea, methodology, and experimental protocols were independently conceived and designed by the authors, with AI tools serving only a supporting role. All AI-assisted contents were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Rameen Abdal, Or Patashnik, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Dynamic concepts personalization from single videos, 2025. URL https://arxiv.org/abs/2502.14844.

Vatsal Baherwani, Yixuan Ren, and Abhinav Shrivastava. Characterizing motion encoding in video diffusion timesteps, 2025. URL https://arxiv.org/abs/2512.22175.

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

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *CVPR*, 2025. Licensed under Modified Apache 2.0 with special crediting requirement.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *ICLR*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*, 2024.

Yinan Chen, Jiangning Zhang, Teng Hu, Yuxiang Zeng, Zhucun Xue, Qingdong He, Chengjie Wang, Yong Liu, Xiaobin Hu, and Shuicheng Yan. IVEBench: Modern benchmark suite for instruction-guided video editing assessment. In *ICLR*, 2026.

Chenjian Gao, Lihe Ding, Xin Cai, Zhanpeng Huang, Zibin Wang, and Tianfan Xue. Controllable first-frame-guided video editing via mask-aware loRA fine-tuning. In *ICLR*, 2026. URL https://openreview.net/forum?id=xkRMJ1Y7Um.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *ICLR*, 2024.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. URL https://arxiv.org/abs/2006.11239.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023. URL https://openreview.net/forum?id=rB6TpjAuSRy.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhi-Lin Huang, Yixuan Liu, Chujun Qin, Zhongdao Wang, Dong Zhou, Dong Li, and Emad Barsoum. Edit as you see: Image-guided video editing via masked motion modeling, 2025. URL https://arxiv.org/abs/2501.04325.

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*, 2024a.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024b.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *ICCV*, 2025.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *ICCV*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *ICCV*, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024.

Jinlai Liu, Jian Han, Bin Yan, Hui Wu, Fengda Zhu, Xing Wang, Yi Jiang, Bingyue Peng, and Zehuan Yuan. InfinityStar: Unified spacetime autoregressive modeling for visual generation. In *NeurIPS*, 2025a.

Shaoteng Liu, Tianyu Wang, Jui-Hsien Wang, Qing Liu, Zhifei Zhang, Joon-Young Lee, Yijun Li, Bei Yu, Zhe Lin, Soo Ye Kim, and Jiaya Jia. Generative video propagation. In *CVPR*, 2025b.

Yue Ma, Yulong Liu, Qiyuan Zhu, Xiangpeng Yang, Kunyu Feng, Xinhua Zhang, Zexuan Yan, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. EffiVMT: Video motion transfer via efficient spatial-temporal decoupled finetuning. In *ICLR*, 2026.

Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. In *ICCVW*, 2025.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Int. Conf. Mach. Learn.*, 2022.

Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. *arXiv preprint arXiv:2405.16537*, 2024.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models. In *ECCV*, 2024.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Yeji Song, Wonsik Shin, Junsoo Lee, Jeesoo Kim, and Nojun Kwak. Save: Protagonist diversification with structure agnostic video editing. In *ECCV*, 2024.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. URL https://arxiv.org/abs/2003.12039.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *NeurIPS*, 2024.

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

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023a.

Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, Rui He, Feng Hu, Junhua Hu, Hai Huang, Hanyu Zhu, Xu Cheng, Jie Tang, Mike Zheng Shou, Kurt Keutzer, and Forrest Iandola. Cvpr 2023 text guided video editing competition, 2023b. URL https://arxiv.org/abs/2310.16003.

Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. In *AAAI*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *ICLR*, 2025. URL https://openreview.net/forum?id=N8Oj1XhtYZ.

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. *ICLR*, 2025.

Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *CVPR*, 2024.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *IJCV*, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. In *ICLR*, 2025. URL https://arxiv.org/abs/2406.07548.

A Implementation Details

A.1 Evaluation Protocol

To ensure a fair and robust quantitative evaluation, we perform inference using two random seeds for each method. From the two generated outputs per video, we average their respective scores to compute the final scores. Regarding the temporal length, ours, I2VEdit (Ouyang et al., 2024), LoRAEdit (Gao et al., 2026), and VACE (Jiang et al., 2025) are trained and evaluated using 81 frames. It is strictly determined by our base pre-trained model, InfinityStar (Liu et al., 2025a), which exclusively supports 81-frame generation. For 80-frame videos, we duplicate one frame to match 81-frame input length. In contrast, Go-with-the-Flow (Burgert et al., 2025) is trained and evaluated using 49 frames, as its publicly released weights are constrained to process a maximum of 49 frames. Therefore, we inevitably have to conduct the evaluation using difference frame lengths. However, for a fair comparison, frames were uniformly sampled from the exact same temporal range across all methods, and all other configurations strictly adhered to the default settings of each respective model. Consequently, the source videos for Go-with-the-Flow are extracted at 49 frames.

Regarding the choice of baselines, we excluded VACE and LoRAEdit from our global editing comparisons. Although attempting global editing with a full spatial mask is technically feasible for these models, their inherent prior lead to generation without preserving the original content. This prevents them from utilizing the model's prior, rendering a direct comparison meaningless. This is why generating without mask is important for first-frame-guided video editing. Additionally, while methods like GenProp Liu et al. (2025b) would serve as excellent baselines, we are unable to include them in our evaluation as their codes and checkpoints are not publicly available.

A.2 Augmentation Strategy

To prevent the Motion LoRA (Hu et al., 2022) from overly relying on the spatial representations of the Appearance LoRA, we introduce a dropout-based augmentation strategy during Stage 2. Specifically, we randomly deactivate the Appearance LoRA and apply spatial augmentations–such as color jittering or Gaussian blur–to the source frames with a probability p. We empirically set p to 0.35 in this paper. By degrading spatial priors and removing Appearance LoRA, it forces the Motion LoRA to learn motion-focused adaptation with reduced reliance on static appearance cues. During these augmentation steps, we apply a 0.5 scaling factor to the BCE loss to reduce its influence, while keeping the regularization weight unchanged. For computational efficiency, we pre-generate ten sets of spatially augmented videos and store their encoded features in CPU memory. During training, these features are dynamically transferred to the GPU when the appearance LoRA is dropped.

A.3 Experimental Details

For our two-stage training process, we utilize distinct LoRA (Hu et al., 2022) hyperparameters for each stage to effectively balance appearance and motion learning. Specifically, in stage 1, we employ a LoRA rank (r) of 16 and an LoRA scaling factor (α_{LoRA}) of 32.0. For stage 2, we increase the capacity to capture complex temporal dynamics by setting the LoRA rank to 32 and the scaling factor to 64.0. All experiments were conducted using a single NVIDIA A100 (80GB) GPU with CUDA version 12.4. The training process required approximately 31GB of VRAM, with execution times of 0.8 seconds per step for stage 1 and 3.8 seconds per step for stage 2. For inference, the memory consumption increased to 66GB due to KV caching, and the generation process takes approximately one minute per video.

B Additional Qualitative Results

In addition to the qualitative results provided in the main paper, we present additional examples. Specifically, Fig. 8 and 9 present results for local editing, while Fig. 10 and 11 illustrate expanded examples of global editing. Further results including videos are provided in the supplementary material to better demonstrate the input consistency and motion fidelity of our approach.

C User Study

To validate the effectiveness of our proposed method, we conduct a user study designed to ensure a fair comparison. We construct two distinct evaluation sets, each consisting of 10 source videos (5 for object and 5 for style), where the generated outputs were randomly selected from the two random seed generations previously used in our quantitative evaluation. To eliminate evaluation bias, all methods are blinded, and the order of the videos is randomized for each test case. Furthermore, we unify the resolution and aspect ratio across all compared methods to maintain visual consistency. To ensure a consistent 81-frame evaluation standard, we apply temporal interpolation to the results of Go-with-the-Flow (Burgert et al., 2025), aligning its frame count with all other methods.

Our human evaluation is conducted with the voluntary participation of 20 individuals who provided informed consent prior to the study, with 10 participants assigned to each set. The study is performed without monetary compensation, and we ensured that no sensitive or personal information is collected during the process. Participants are asked to rank the generated videos for every given source. The final results are reported as the average rank, where a lower score indicates higher quality and human preference. We find that some participants provided unreliable or missing responses for certain cases; these responses were excluded from the final aggregation. A sample interface of our user study is presented in Fig. 6 and 7.

Table 5: Ablation study on the effect of two-stage training.

Method	**Visual Quality**		**Input Consistency**			**Temporal Consistency**
	Imaging	Aesthetic	Subject	Background	CLIP-I2V	Motion Fidelity
Joint Training	**0.6537**	**0.5887**	**0.9732**	**0.9656**	0.8838	0.7100
Ours	0.6401	0.5845	0.9717	0.9643	**0.8862**	**0.7533**

Table 6: Ablation study on the effect of the Appearance scale cutoff.

Exclusive Cutoff (k_c)	**Visual Quality**		**Input Consistency**			**Temporal Consistency**
	Imaging	Aesthetic	Subject	Background	CLIP-I2V	Motion Fidelity
6 (Ours)	0.6401	0.5845	0.9717	**0.9643**	0.8862	**0.7533**
9	**0.6419**	**0.5880**	**0.9730**	0.9638	**0.8863**	0.7473
12	0.6358	0.5850	0.9710	0.9632	0.8846	0.7454
All	0.6365	0.5846	0.9711	0.9630	0.8820	0.7487

Table 7: Ablation study on the effect of flow-guided regularization weight.

Method	**Visual Quality**		**Input Consistency**			**Temporal Consistency**
	Imaging	Aesthetic	Subject	Background	CLIP-I2V	Motion Fidelity
$\boldsymbol{\alpha = 0.001}$	0.6560	0.5892	0.9726	**0.9644**	0.8880	**0.7667**
$\alpha = 0.005$	0.6572	**0.5916**	0.9739	0.9641	0.8867	0.7587
$\alpha = 0.010$	**0.6588**	0.5905	**0.9746**	0.9639	0.8874	0.7571
$\alpha = 0.050$	0.6570	0.5902	0.9740	0.9643	**0.8894**	0.7559

D Ablation Study

Owing to the computational cost, ablation studies in this section are conducted on subsets randomly sampled from our full evaluation set except Sec. D.3. Sec. D.1–D.2 share the same evaluation subset of 40 videos and Sec. D.3 is conducted on the same dataset in main evaluation, whereas D.4 is conducted as a separate cutoff sweep in which each variant is independently trained and regenerated on 20 videos.

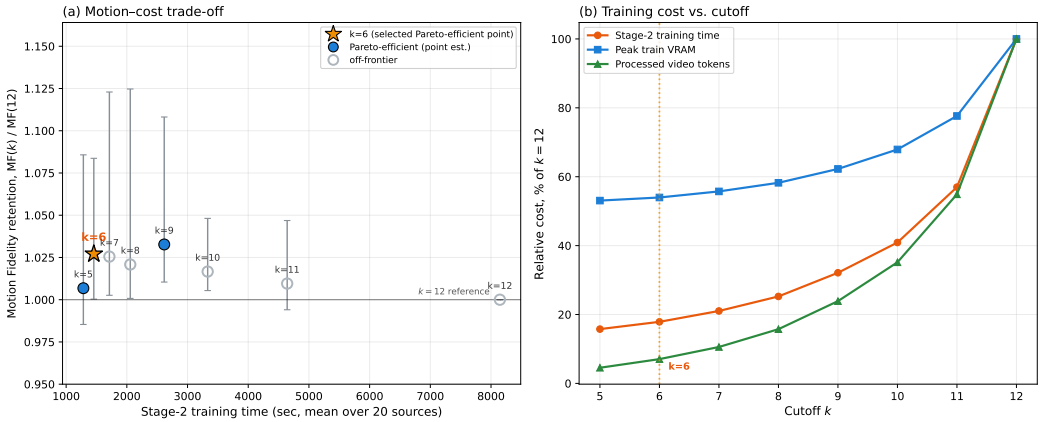

Figure 5: **Ablation study on extending training to higher scales.** (a) Motion Fidelity relative to the largest feasible cutoff training under the fixed 400-step schedule. (b) Stage-2 training time, peak training memory, and processed video tokens, normalized by their respective values at $k = 12$. The cutoff $k = 6$, selected using the 95% reconstruction criterion, achieves a favorable balance between motion preservation and adaptation cost.

D.1 Ablation Study on Joint Training

Building upon the evaluations in the main paper, we conduct additional ablation studies to further justify our design choices. Specifically, we investigate a "joint training" baseline where the Appearance LoRA (Hu et al., 2022) and Motion LoRA (Hu et al., 2022) are optimized simultaneously, in contrast to our proposed two-stage strategy. As shown in Tab. 5, the results indicate that both methods achieve comparable performance in terms of visual quality and input consistency. However, they exhibit a critical gap in motion alignment. Due to suboptimal optimization caused by entangled learning objectives, the joint training scheme severely fails to follow the motion of source videos. Consequently, the model fails to capture the temporal dynamics of the source video and instead produces simplistic motion patterns inherent in the prior knowledge of the base model. While its spatial metrics remain comparable to ours, the joint training baseline yields an exceptionally low motion fidelity score, quantitatively demonstrating its inability to capture and transfer the temporal dynamics of the source videos.

D.2 Ablation Study on the Appearance Scale

We further investigate the impact of the target image scale k utilized during the training of the Appearance LoRA (Hu et al., 2022) in stage 1. As demonstrated in Tab. 6, expanding the image scale beyond the target scale required for the subsequent video training phase yields comparable performance in overall metrics. Given the multi-scale nature of the VAR (Tian et al., 2024; Liu et al., 2025a) framework where the number of tokens expands significantly at higher scales, increasing k incurs substantial computational overhead, drastically degrading training speed. Consequently, we strictly align the target image scale as $k \in \{0, 1, ..., 5\}$ with the video scales. It enables stage 1 to achieve exceptionally fast and memory-efficient training without compromising the performance.

D.3 Ablation Study on the Flow-Guided Regularization Weight

We conduct an ablation study to determine the optimal weight α for our flow-guided regularization. Our experiments reveal that an excessive regularization disrupts the balance with the BCE Loss, hindering the generative process. As shown in Table 7, visual quality and input consistency remain comparable across different coefficients. Therefore, we adopt motion fidelity as the primary criterion for hyperparameter selection. The results indicate that $\alpha = 0.001$ yields the optimal balance, achieving the highest motion fidelity.

D.4 Ablation Study on Extending Training to Higher Scales

Motivated by the analysis in Sec. 3.2, we select the exclusive cutoff $k_c = 6$, for which $S_6/S_{recon} = 0.951$, retaining over 95% of the recoverable source motion. This indicates that the first six scales contain sufficient information to reconstruct most of the source motion. Nevertheless, extending training to higher scales may provide additional performance gains, motivating a further analysis of the cutoff choice.

For this analysis, we independently train two-stage LoRA (Hu et al., 2022) for each cutoff $k \in \{5, \ldots, 12\}$. Since it is computationally intensive, we conduct the experiment on 20 source videos randomly sampled from our full evaluation set. All variants are trained using the same 400-step schedule and identical optimization hyperparameters. This fixed-update protocol follows our standard adaptation setting and allows us to evaluate whether extending the cutoff is justified when both motion fidelity and practical adaptation efficiency are considered. Accordingly, this experiment evaluates cost-effectiveness under a common adaptation budget rather than the fully converged performance ceiling of each cutoff.

As k increases, the spatiotemporal grids of the additional scales become substantially larger, causing the number of processed visual tokens to rise sharply. Consequently, higher cutoffs sharply increase the number of processed tokens, peak memory consumption, and training time per step. Although higher cutoffs may potentially benefit from additional optimization steps, increasing the training schedule would further amplify their already substantial per-step computational cost.

Under this controlled setting, extending the cutoff beyond $k = 6$ provides no consistent improvement in Motion Fidelity while substantially increasing training cost. When training cost is minimized and Motion Fidelity is maximized, $k = 6$ lies on the point-estimate Pareto frontier under the fixed-step protocol: no evaluated cutoff achieves higher median Motion Fidelity at a lower training cost. Since the confidence intervals overlap, we interpret this result as evidence of cost-effectiveness rather than statistical dominance over every larger cutoff. Combined with the independent 95% reconstruction criterion from Sec. 3.2, this supports $k = 6$ as a cost-effective operating point, rather than claiming it as a universally optimal cutoff. We exclude $k = 13, 14$ since training at these cutoffs exceeds the available GPU memory. An overview is provided in Fig. 5.

E Limitations

While our method is highly effective for first-frame guided video editing, it possesses some limitations. Most notably, the performance of our framework is fundamentally constrained by the inherent capabilities of the underlying base model, InfinityStar (Liu et al., 2025a). Specifically, the model is strictly limited to generating a single video clip of exactly 81 frames at 720p resolution, lacking the flexibility for arbitrary frame counts of long-form generation. Furthermore, its visual quality inherently lags behind state-of-the-art diffusion models that have been refined over years of extensive data curation and engineering optimization. Consequently, while our method excels at following precise motion dynamics, the generated videos occasionally exhibit visual degradations, such as blurriness or loss of fine-grained details. Notably, similar degradations also occur in vanilla InfinityStar. Our pipeline relies on the frozen base model to generate the remaining high-resolution scale maps that refine appearance details, and empirical observations confirm that identical detail collapses consistently manifest during vanilla inference using the base model alone. We anticipate that these visual limitations will be naturally resolved as the underlying Text-to-Video VAR architectures continue to evolve.

F Societal Impact

Our VAR-based video editing framework provides an efficient tool for creative expression and content production. While it significantly reduces manual labor for artists and researchers, we acknowledge that, like existing image and video editing technologies, it carries the inherent risk of generating misleading synthetic content (e.g. DeepFakes). To mitigate these concerns, we strictly utilized publicly available datasets for all experiments and evaluations reported in this paper. We believe that acknowledging these risks is an essential step toward the responsible development of video synthesis.

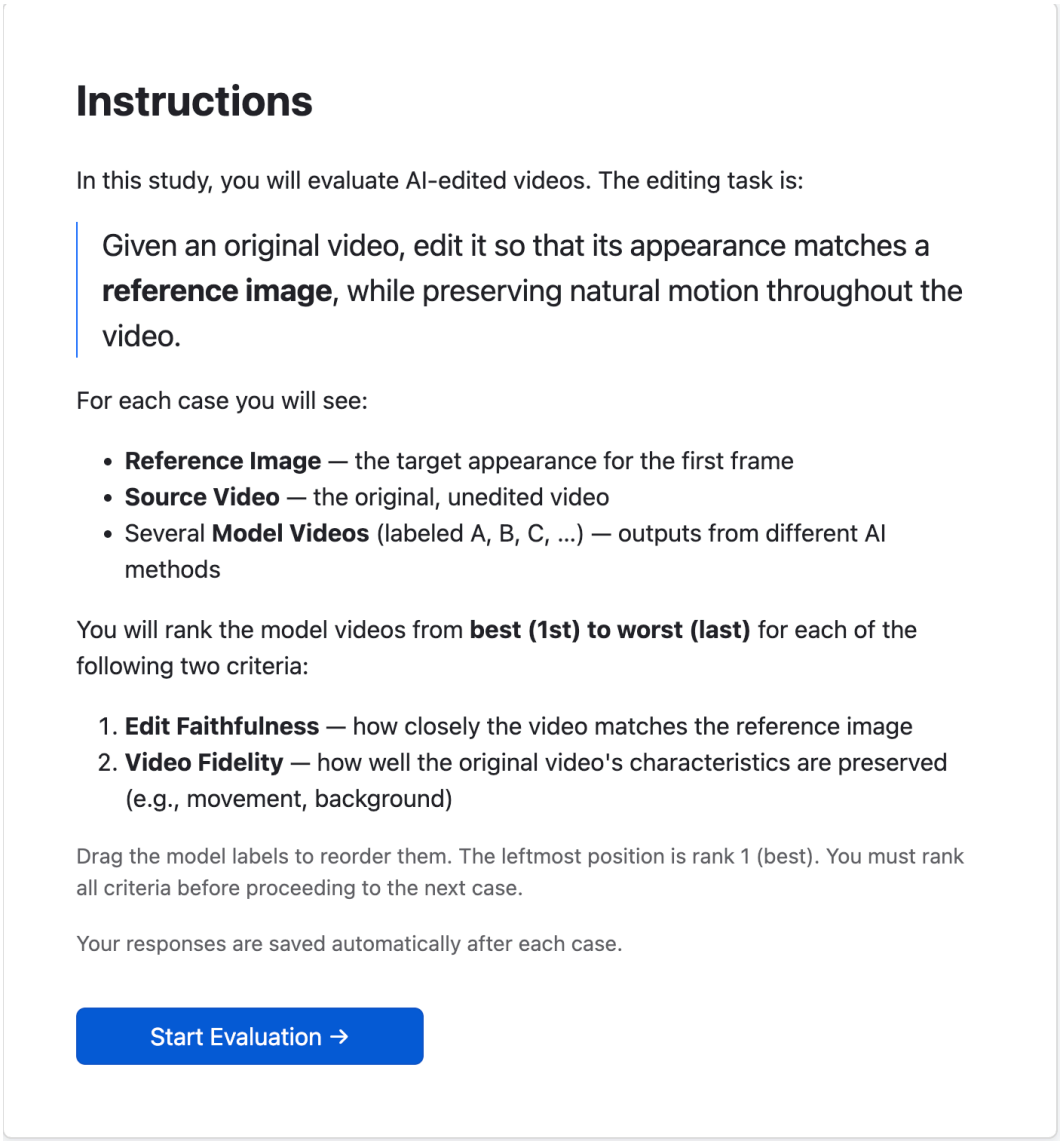

Figure 6: **User Study Sample.**



Figure 7: **User Study Sample.**

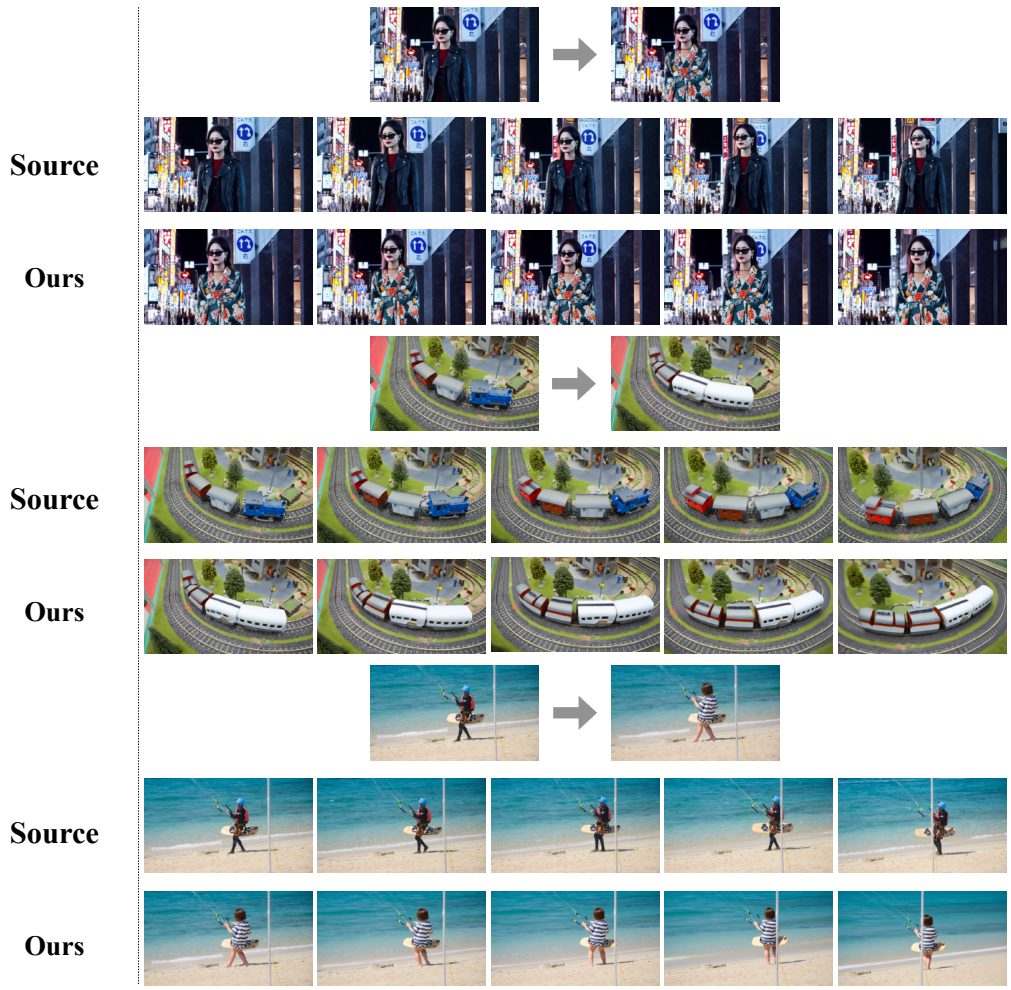

Figure 8: **Additional Qualitative Results.**



Figure 9: **Additional Qualitative Results.**

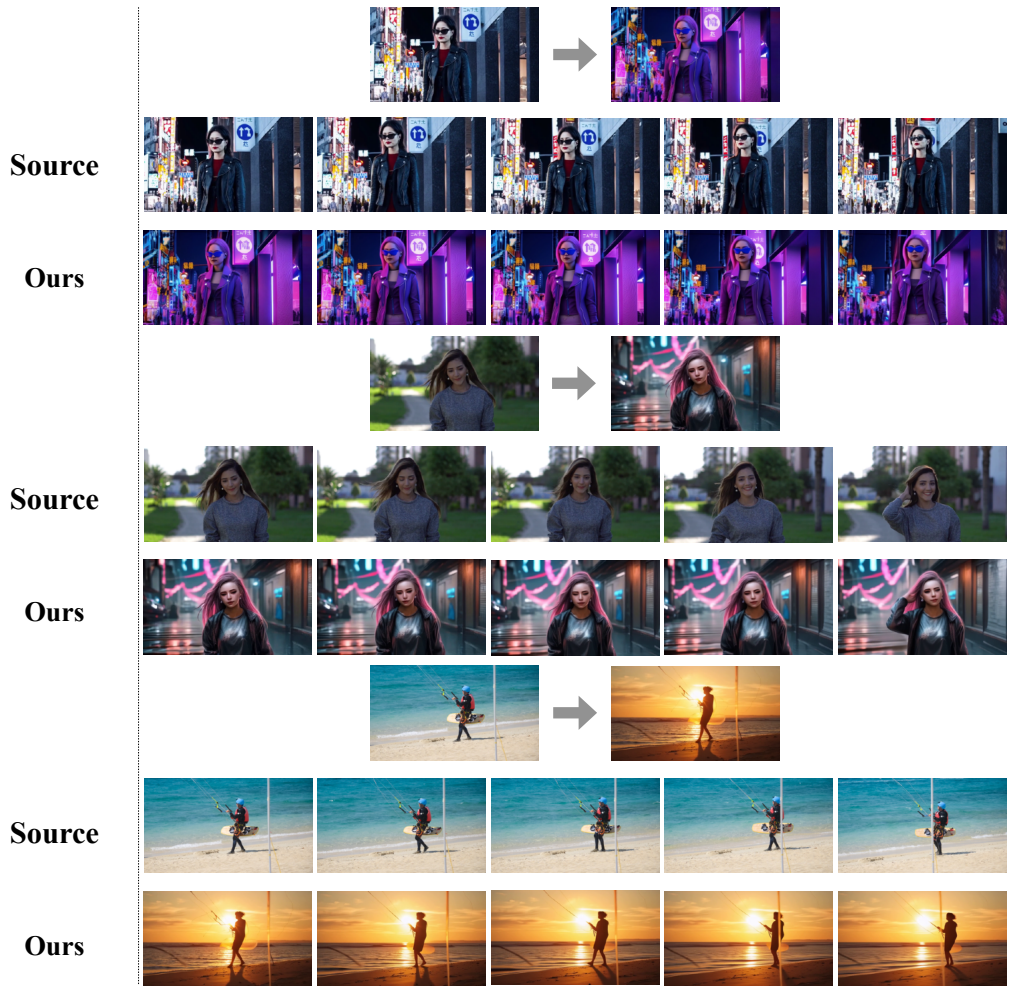

Figure 10: **Additional Qualitative Results.**



Figure 11: **Additional Qualitative Results.**