VIA: A Spatiotemporal Video Adaptation Framework for Global and Local Video Editing

Jing Gu[1] **Yuwei Fang**[2] **Ivan Skorokhodov**[2] **Peter Wonka**[3] **Xinya Du**[4]
Sergey Tulyakov[2] **Xin Eric Wang**[1]
[1]University of California, Santa Cruz [2]Snap Research
[3]KAUST [4]University of Texas at Dallas
{jgu110, xwang366}@ucsc.edu, yfang3@snapchat.com
https://via-video.github.io/

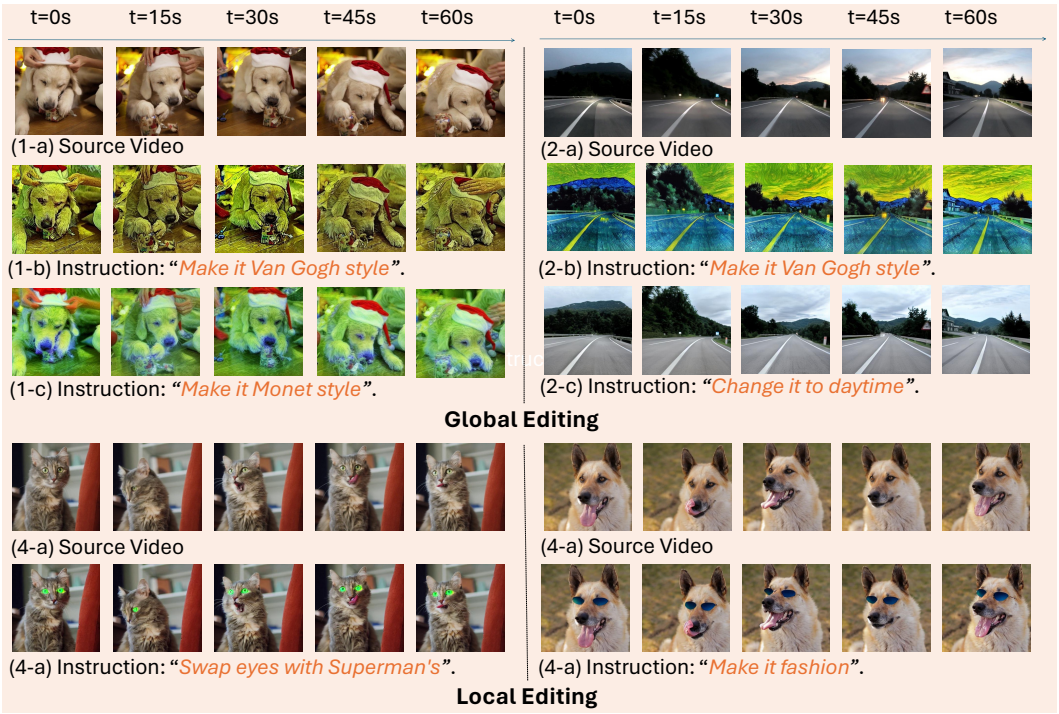


Figure 1: **Video editing results by VIA.** VIA excels in *precise* and *consistent* editing across diverse video editing tasks. Above show consistent results over long videos with duration of 1 minute, which is challenging in current literature. Below show consistent results of precise local editing.

Abstract

Video editing stands as a cornerstone of digital media, from entertainment and education to professional communication. However, previous methods often overlook the necessity of comprehensively understanding both global and local contexts, leading to inaccurate and inconsistency edits in the spatiotemporal dimension, especially for long videos. In this paper, we introduce VIA, a unified spatiotemporal VIdeo Adaptation framework for global and local video editing, pushing the limits of consistently editing minute-long videos. First, to ensure local consistency within individual frames, the foundation of VIA is a novel *test-time editing adaptation* method, which adapts a pre-trained image editing model for improving consistency

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

between potential editing directions and the text instruction, and adapts masked latent variables for precise local control. Furthermore, to maintain global consistency over the video sequence, we introduce *spatiotemporal adaptation* that adapts consistent attention variables in key frames and strategically applies them across the whole sequence to realize the editing effects. Extensive experiments demonstrate that, compared to baseline methods, our VIA approach produces edits that are more faithful to the source videos, more coherent in the spatiotemporal context, and more precise in local control. More importantly, we show that VIA can achieve consistent long video editing in minutes, unlocking the potentials for advanced video editing tasks over long video sequences.

1 Introduction

With the exponential growth of digital content creation, video editing has become indispensable for various purposes, including filmmaking [1, 2], advertising [3, 4], education [5, 6], and social media [7, 8]. This task presents significant challenges, primarily in preserving the integrity of the original video, accurately executing user instructions, and maintaining consistent editing quality over time and space. These challenges become particularly pronounced with longer videos, where ensuring long-range spatiotemporal consistency is critical.

To tackle these challenges, an extensive range of methods have been proposed [9–12]. One promising approach, building on the remarkable success of image-based diffusion models [13–17], is to adapt their image editing capabilities to ensure temporal consistency during test time [18, 19, 12, 20, 21]. However, there remains a crucial oversight in comprehensively addressing both global and local contexts, leading to inaccuracies and inconsistencies in edits across spatiotemporal dimension. Most existing methods are constrained by their ability to maintain spatiotemporal consistency, typically limiting their edits to video sequences only seconds in duration. In fact, the source video itself can be the best guide due to its inherent spatiotemporal consistency. Therefore, achieving effective video editing across longer durations requires novel approaches that not only ensure temporal consistency during testing but also encompass a deeper understanding of source video cues at both global and local levels.

In this work, we introduce VIA, a unified spatiotemporal video adaptation framework towards faithful, consistent, and precise video editing, pushing the limits of editing minute-long videos. First, the foundation of our work is a novel *test-time editing adaptation* that adapts a pretrained image editing model, to improve semantic understanding of the source video and consistency between potential editing directions and the text instruction. We propose an augmentation pipeline to obtain the in-domain tuning set for test-time adaptation, where the image editing model learns to associate specific visual editing directions with the provided instructions, significantly enhancing semantic understanding and editing consistency within individual frames. To further enhance local consistency, we introduce local latent adaptation with automated mask generation powered by the multimodal large language model and segmentation model, which achieves precise local control of the editing targets across frames.

Second, adapting image editing models inherently lacks spatiotemporal consistency when applied to video frames. To address this, we introduce *spatio-temporal attention adaptation* for maintaining global editing consistency across frames. Specifically, we propose a gather-and-swap strategy for efficient global editing, which leverages consistent attention variables from the model's architecture, and strategically applies them across the sequence. This method not only aligns with the continuity of the video but also reinforces the fidelity of the editing process, ensuring that changes are harmonized across frames and over time.

Through rigorous testing and evaluation, our methods have demonstrated superior performance over existing techniques, offering significant improvements in both local edit precision and the overall aesthetic quality of videos. By pushing the boundaries of what is possible in video editing, our work opens new avenues for media production and creative content generation, marking a significant step forward in the integration of AI-driven techniques into video editing. To the best of our knowledge, we are the first to achieve minutes-long video editing.

2 Related Work

2.1 Text-driven Video Editing

Text-driven Video Editing is a process to modify videos according to the instructions given by user. Inspired by the remarkable success of text-driven image editing [16, 17, 22–24], extensive methods have been proposed for video content editing [25, 18, 19, 12, 20, 21, 26]. One paradigm for video editing is to adapt an image-based model to video. For example, [18] adapts image editing to the video domain without any training or fine-tuning by changing the self-attention mechanisms in Instruct-Pix2Pix to cross-frame attentions. [19] explicitly propagates diffusion features based on inter-frame correspondences to enforce consistency in the diffusion feature space. [27] construct a neural video field to enable encoding long videos with hundreds of frames in a memory-efficient manner and then update the video field with image-based model to impart text-driven editing effects. [26] plug in any existing image editing tools to support an extensive array of video editing tasks. However, these methods are constrained by their ability to maintain global and local consistency, limiting to edit short videos within seconds. To efficiently enable longer video editing, [12] centers on the concept of anchor-based cross-frame attention, firstly achieving editing 27 seconds videos. In our work, we built upon this line of work and improve editing and spatiotemporal consistency, firstly pushing the limits of video editing to minutes-long videos.

2.2 Test-time Adaptation

Image-based video editing faces the challenge of ensuring temporal consistency during test time. To address this, [10] propose to finetune a text-to-image model on a test video, enabling the generated videos with similar motion patterns to the source video. [9] proposes light-weight spatial and temporal adapters for efficient one-shot video editing. [11] adds a motion modeling module to the frozen based text-to-image model, and trains it on video clips, thereby distilling a reasonable motion prior. [12] uses the same training set that was used to training the image editing model, and applies a data augmentation strategy for continuing pretraining to make the model equivariant to affine transformations. Different from the above approaches, we propose two orthogonal approaches that employs inference-time finetuing and local latent adaption, ensuring consistent and precise editing across frames.

2.3 Spatiotemporal Consistency

Ensuring spatiotemporal consistency is critical for video editing, especially for long videos. [20] makes the attempt to study and utilize the cross-attention and spatial-temporal self-attention during DDIM inversion. [21] proposes a spatial regularization module to fidelity to the original video. [28] presents spectral motion alignment (SMA), a framework that learns motion patterns by incorporating frequency-domain regularization, facilitating the learning of whole-frame global motion dynamics, and mitigating spatial artifacts. [12] improves the design of spatial temporal attention to anchor-based cross-frame attention to ensure spatiotemporal consistency. In our work, we further ensure consistency inside the anchor-based frames and propose a two-step gather-swap process to adapt spatiotemporal attention for consistent global editing.

3 Preliminaries

Diffusion Models In this work, we adapt existing image editing model for instruction-based video editing. Given an image x, the diffusion process produces a noisy latent $\boldsymbol{z}_t$ from the encoded latent $z = \mathcal{E}(x)$ where the noise level increases over timesteps $t \in T$. A network $\boldsymbol{\epsilon}_\theta$ is trained to minimize the following optimization problem,

$$\min_{\theta} \mathbb{E}_{y,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(c_I), c_T)\right\|\right] \tag{1}$$

where $\epsilon \in \mathcal{N}(0,1)$ is the noise added by the diffusion process and $y = (c_T, c_I, x)$ is a triplet of instruction, input image and target image. Here ϵ_θ usually operate on the U-Net architecture [29], including convolutional blocks, as well as self-attention and cross-attention layers.

Attention Layer The attention layer first computes the attention map using query, $\mathbf{Q} \in \mathbb{R}^{n_q \times d}$, and key, $\mathbf{K} \in \mathbb{R}^{n_k \times d}$ where d, n_q and n_k are the hidden dimension, the numbers of the query and key

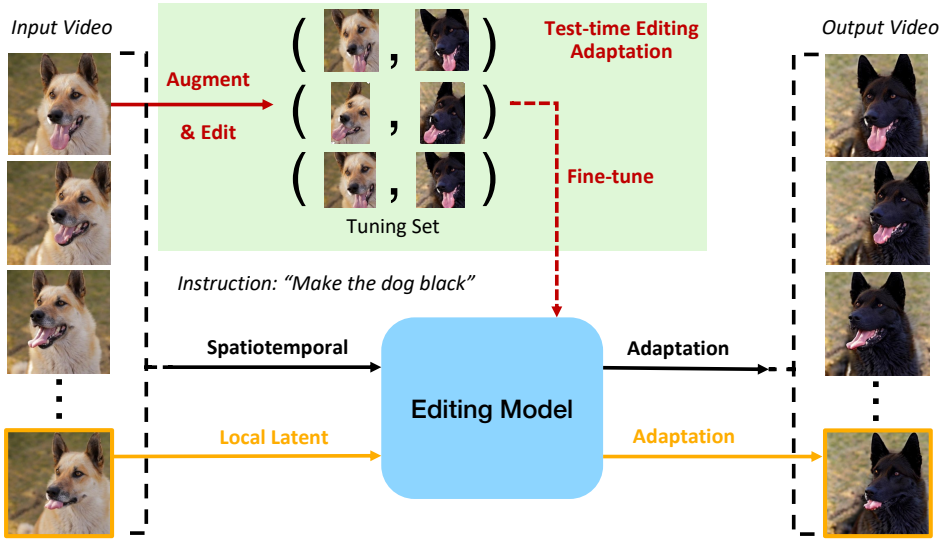


Figure 2: **Overview of our VIA framework.** For local consistency, Test-time Editing Adaptation finetunes the editing model with augmented editing pairs to ensure the consistent editing directions with the text instruction, and Local Latent Adaptation achieves precise editing control and preserves non-target pixels from the input video. For global consistency, Spatiotemporal Adaptation collects and applies key attention variables across all frames.

tokens respectively. Then, the calculated attention map is applied to the value, $\mathbf{V} \in \mathbb{R}^{n \times d}$, describing as follows:

$$\mathbf{Z}' = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V}, \tag{2}$$

$$\mathbf{Q} = \mathbf{Z}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{C}\mathbf{W}_k, \quad \mathbf{V} = \mathbf{C}\mathbf{W}_v, \tag{3}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are the projection matrices to map the different inputs to the same hidden dimension d. $\mathbf{Z}$ is the hidden state and $\mathbf{C}$ is the condition. For self attention layers, the condition is the hidden state while the condition is text conditioning in cross attention layers.

Cross-frame Attention Given N frames from source video, cross-frame attention has been employed in video editing by incorporating $\mathbf{K}$ and $\mathbf{V}$ from previous frames into the current frame's editing process [30, 21, 12], as shown below:

$$\phi = \text{Softmax}\left(\frac{\mathbf{Q}_{\text{curr}}[\mathbf{K}_{\text{curr}}, \mathbf{K}_{\text{group}}]^{\mathbf{T}}}{\sqrt{d}}\right)[\mathbf{V}_{\text{curr}}, \mathbf{V}_{\text{group}}], \tag{4}$$

where $\mathbf{K}_{\text{group}} = [\mathbf{K}^0, \ldots, \mathbf{K}^k]$ and $\mathbf{V}_{\text{group}} = [\mathbf{V}^0, \ldots, \mathbf{V}^k]$, and k is the group size. By incorporating $\mathbf{K}_{\text{group}}$ and $\mathbf{V}_{\text{group}}$ during the video editing process for each frame, the temporal consistency is improved. In this paper, we improve cross-frame attention with a two stage gather-swap process to significantly improve the spatiotemporal consistency.

4 The VIA Framework

We introduce a unified framework to address key challenges in instruction-guided video editing, particularly the issues of editing consistency and spatiotemporal consistency across video frames using an image editing model as in Fig. 2. Below, we detail the distinct methodologies that underpin our approach, each tailored to solve specific aspects of video editing challenges.

4.1 Test-Time Editing Adaptation for Consistent Local Editing

Videos typically exhibit significant variances across the temporal dimension, especially for long videos. When adapting image editing models for video editing, the same edit instructions can lead to

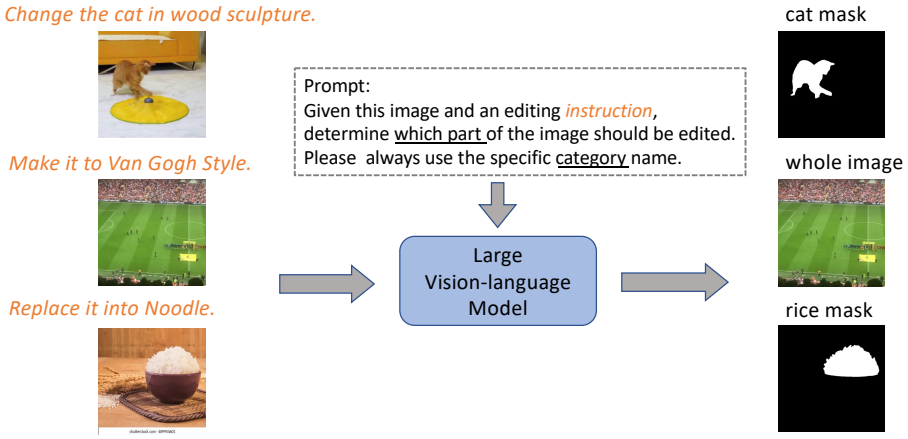


Figure 3: **Automatic mask generation.** A single frame from the video, along with a tailored text prompt that encapsulates the editing instruction, is fed into a Large Vision-Language Model (LVLM), such as GPT-4, to generate a text description specifying the area to be edited. If the designated editing area does not encompass the entire image, this text description is then input into a segmentation model to create a mask for the targeted area.

diverse semantic interpretations and indicate different areas that should be altered in different frames. This inherent variability makes it challenging to maintain consistent edits throughout the entire video. To this end, we propose two orthogonal approaches for consistent local editing.

Drawing inspiration from DreamBooth [31], which employs inference-time fine-tuning to associate specific objects with unique textual tokens, we also link visual editing outcomes with corresponding editing instructions. We first propose a pipeline to obtain the in-domain tuning set without needing external resources. For the video to be edited, the image editing model Ψ first edit a randomly sampled frame S_{root} with different random seed, and choose the best editing result E_{root}. Given the choosing editing result, E_{root}, we use random affine transformation to both edited frame and source frame. Consider $\mathcal{F}_k$ as the affine transformation:

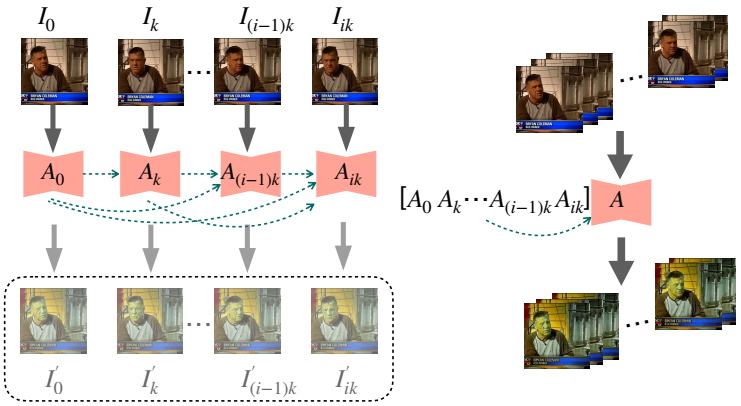
$$T = \{(\mathcal{F}_k(S), \mathcal{F}_k(E), I) \mid \mathcal{F}_k \in \mathcal{F}\} \tag{5}$$

where $\mathcal{F}$ is the set of transformation. So the tuning set T, consists of triples: source image, edited image, and editing instruction. By fine-tuning the image editing model Ψ on this domain-specific dataset, the model learns to associate specific visual editing directions with the provided instructions.

This approach offers two significant benefits. Firstly, it enhances semantic consistency across the video, particularly for instructions that lack detailed editing specifications, by mitigating divergent editing outcomes for different frames. Secondly, it improves the overall quality of edits by using the best editing outcome as the root pair for further fine-tuning, a method particularly effective in localized edits where performance might otherwise be less robust. In this way, the image editing model reinforces its most successful editing strategies.

Local Latent Adaptation Editing instructions may indicate that only a particular area should be altered. However, current end-to-end models for following such instructions often inadvertently modify areas not targeted by the user. To address this issue, we propose a method for precise editing. Initially, a Large Vision-Language Model (LVLM) is prompted to provide a textual description, P, of the area to be swapped for each frame. Based on this description, P, we employ the Segment Anything model [32] to extract a mask delineating the area to be edited.

Previous methods in image domain achieved localized editing via directly blending the latent $\boldsymbol{z}$ from source image and target image at each diffusion step [33, 34]. For video editing, rather than just the source frame, the editing process should also consider other frames and maintain in-frame consistency. Here, we propose to anneally increase the blend the latent from source frame to target frame. During the diffusion process, we merge the inverted latent representation with the generated latent at each timestep. Furthermore, we observed that this blending process benefits the overall video editing workflow by ensuring that edits are confined to the targeted areas. By maintaining the integrity of non-targeted regions, our approach compels the model to adhere strictly to the editing instructions and focus on the specified areas. We further propose **Progressive Boundary Integration** to smoothly



(a) Gather attention group (b) Edit all frames with attention group.

Figure 4: **The gather and swap process for video editing.** The left part of the diagram illustrates the gathering process. We initially select $k+1$ frames evenly distributed throughout the video. The first frame undergoes standard editing using an image editing model, during which the attention variables are captured and stored. For each of the subsequent k frames, the attention variable from the preceding frame is swapped in, and its own attention variables are also preserved. In the right part, the collected attention variables from all $k+1$ frames are swapped into the editing process of each frame. This includes applying the previously gathered attention variables to enhance the consistency and quality of edits across the sequence.

merge the source latent and the target latent. This is accomplished through a linear interpolation between the values 0 and 1 across the series. The mathematical representation is given by:

$$\mathbf{M}_{\text{src}}(x,y) = \begin{cases} \mathbf{M}_{\text{src}}(x,y) \cdot \frac{t}{T}, & \text{if } t \leq T \text{ and } \mathbf{M}_{\text{src}}(x,y) = 1 \\ \mathbf{M}_{\text{src}}(x,y), & \text{otherwise} \end{cases} \tag{6}$$

Here, $\mathbf{M}_{\text{src}}(x,y)$ is predefined as 1 in a specific central area and 0 elsewhere. Within this central area, $\mathbf{M}_{\text{src}}(x,y)$ incrementally increases from 0 to 1 over T steps, while the values outside this central region remain unchanged.

4.2 Spatiotemporal Adaptation for Consistent Global Editing

To improve the inefficiency issue of cross-frame attention, [12] proposed to sample a set of group frames and edit them with an image-based model. However, this approach does not ensure consistency inside the group frames as the attention variables in this group are still generated independently. Thus, we propose a two-step gather-swap process to adapt spatiotemporal attention for consistent global editing.

Firstly, in the group gathering stage, the model progressively edits the image, with key $\mathbf{K}$ and value $\mathbf{V}$ from previous frames in the group, rather from their own $\mathbf{K}_{\text{curr}}$ and $\mathbf{V}_{\text{curr}}$,

$$\phi = \text{softmax}\left(\frac{\mathbf{Q}_{\text{curr}}\mathbf{K}_{\text{prev}}^T}{\sqrt{d}}\right)\mathbf{V}_{\text{prev}}, \tag{7}$$

$$\mathbf{K}_{\text{group}}^{(t+1)} = [\mathbf{K}_{\text{group}}^{(t)}, \mathbf{K}_{\text{curr}}], \quad \mathbf{V}_{\text{group}}^{(t+1)} = [\mathbf{V}_{\text{group}}^{(t)}, \mathbf{V}_{\text{curr}}] \tag{8}$$

Since $\mathbf{K}_{\text{curr}}$ and $\mathbf{V}_{\text{curr}}$ are calculated by the ϕ from the last layer, which already has a stronger dependency on other frames, the saved elements have a stronger consistency with previous group elements, leading to in-group consistency.

In the second stage, we utilize the attention group in the editing process of all frames, including the frames used to generate the attention group. In this way, we address the inconsistency of the first few frames in the group, where they have less dependency on other frames. During the editing process, each frame still does not utilize its own attention variables:

$$\phi = \text{softmax}\left(\frac{\mathbf{Q}_{\text{curr}}\mathbf{K}_{\text{group}}^T}{\sqrt{d}}\right)\mathbf{V}_{\text{group}}, \tag{9}$$

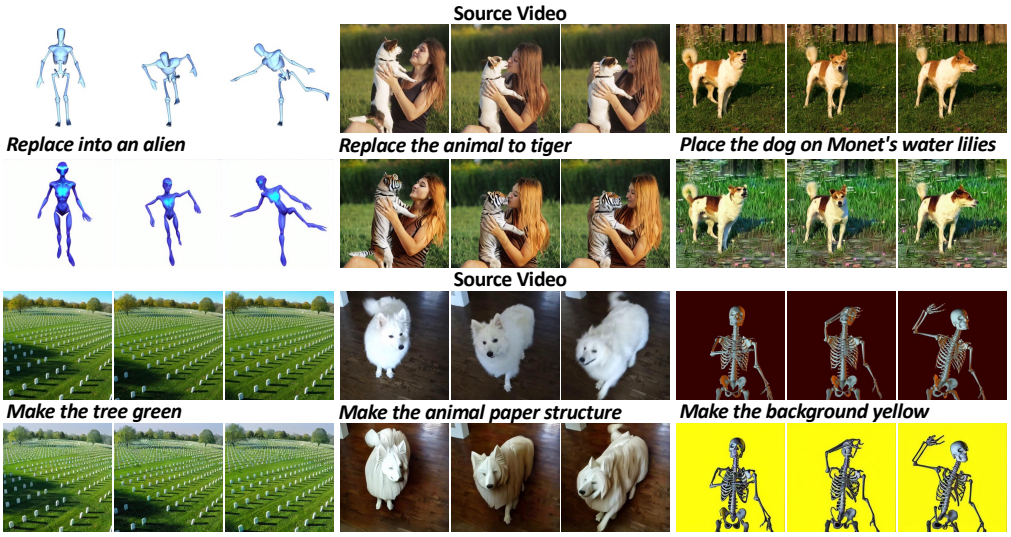


Figure 5: **Local editing results.** VIA is capable of diverse local editing tasks where only part of the pixels in the frame should be altered, including identity swapping, object part editing, background editing.

in this way, all frames share the same attention group that is consistent in itself, which yields maximum consistency between edited frames. Furthermore, previous work has only utilized self-attention for cross-frame attention. We found that cross-attention also serves a similar function, and combining both yields maximum editing results. To maximize coverage of the dynamic changes within a video, we select attention variables from frames that are evenly distributed throughout the video, thereby ensuring a broad representation of frame differences. Figure 4 illustrates the two stages, where $\mathbf{A}$ represents both $\mathbf{K}$ and $\mathbf{V}$.

5 Evaluation

In this paper, we adapt open source image editing model MGIE [35] for video editing. For spatiotemporal adaptation, we gather attention variables from 4 frames. For editing adaptation, we design folowing transformations for each image-pair to enhance variability while maintaining the structural integrity of the images: (i) slight rotation (up to ±5 degrees); (ii) translation (up to 5% both horizontally and vertically); (iii) following the previous two, cropping transformed images from 75% to 100% of its original size to simulate changes in the framing of video sequences. Additionally, the images are sheared by up to 10 degrees. These affine transformations introduce realistic variations to mimic the diversity of viewing angles typically encountered across different frames in a video. We prompt GPT-4V to provide the localized textual descrption and extract editing mask using Grounding DINO [36] and Segment Anything [37].

For comprehensive evaluations with state-of-the-art work, we first compare our results with the closed source method, Fairy [12], which handles longest videos up to 27 seconds in length, using the video from their paper. We then conduct qualitative and human evaluations with open-sourced state-of-the-art baselines, AnyV2V [26], Rerender [38], Tokenflow [19], Video-P2P [30], and Tune-A-Video [10]. For AnyV2V, we use the first edited frame from VIA.

5.1 Long Video Editing

A direct consequence of the high consistency feature in our video editing framework is its proficiency in editing longer videos. While one of our baselines, Fairy [12], has not made their code publicly available, they report that their model can handle videos up to 27 seconds in length. We compare our results on the same video using sample identical instructions in Figure 8. Remarkably, VIA demonstrate superior global and local consistency, attributable to our unified adaptation framework.

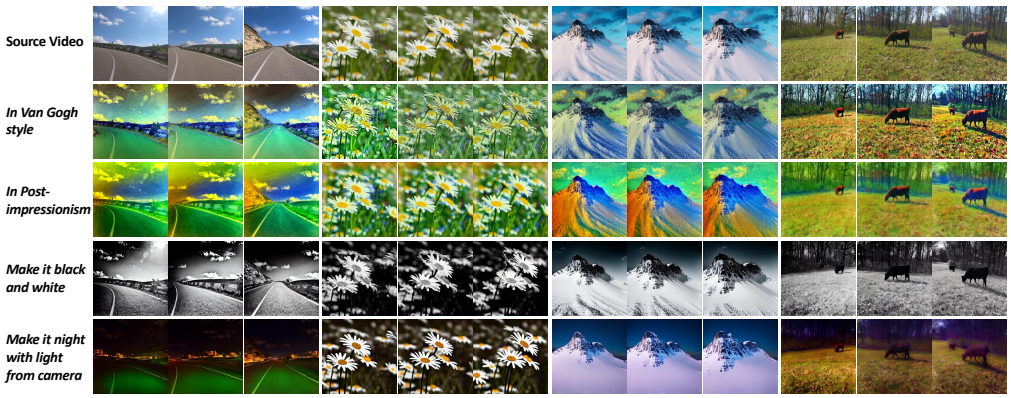


Figure 6: **Global editing results.** VIA demonstrates robust global editing performance across various videos using a consistent set of editing instructions, producing high-quality results.

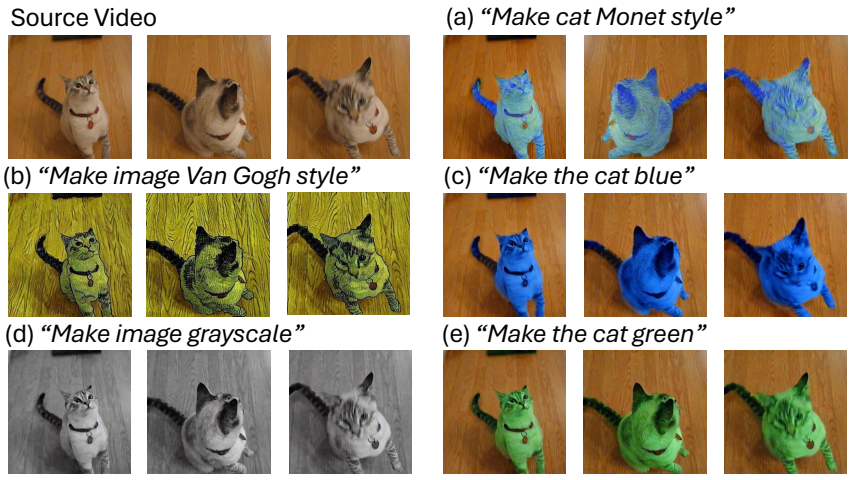


Figure 7: **Global and local stylization.** We show video editing results with different given instructions in (a)-(e). Local Editing in VIA is not limited to object swapping. Whereas other methods can only do stylization on the whole image, our model could achieve a local stylization.

5.2 Qualitative Results

Local editing results. Fig. 5 demonstrates the performance of VIA on various local editing tasks where only specific parts of the frame are altered. VIA accurately identifies the target position and performs precise edits, even with occluded subjects, as seen in the "Replace the animal with a tiger" example. In addition to editing foreground subjects, VIA excels at background modifications. For instance, it can seamlessly "Place the dog on Monet's water lilies" in a video. In the challenging skeleton video, where the background needs to fill the gaps between the bones, VIA maintains consistent performance without affecting the dancing skeleton.

Global editing results. In Fig. 6, we show the global editing performance of VIA across various videos using a consistent set of editing instructions, yielding high-quality results. The same set of editing prompts was applied to different videos. The bottom example illustrates VIA's ability to understand and consistently apply visual effects across all frames.

Fig. 7 demonstrates the advanced video editing capabilities of our method, highlighting its ability to perform both global and local stylization. Unlike previous methods, which are limited to applying stylistic changes to the entire image, our approach allows for precise, localized edits. This flexibility is illustrated through various examples in subfigures (a)-(e), where different instructions are applied to achieve distinct editing effects. Whether it's object swapping or specific regional stylization, our model surpasses the limitations of traditional methods by enabling targeted modifications while preserving the overall composition and aesthetic integrity of the video.

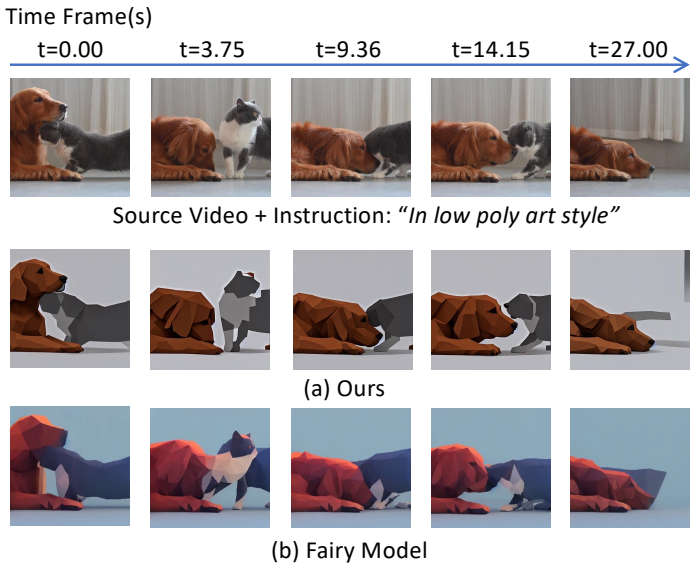


Figure 8: **Comparison with the baseline model on the long video.** We visualize editing results on sampled frames from a 27-second duration video.

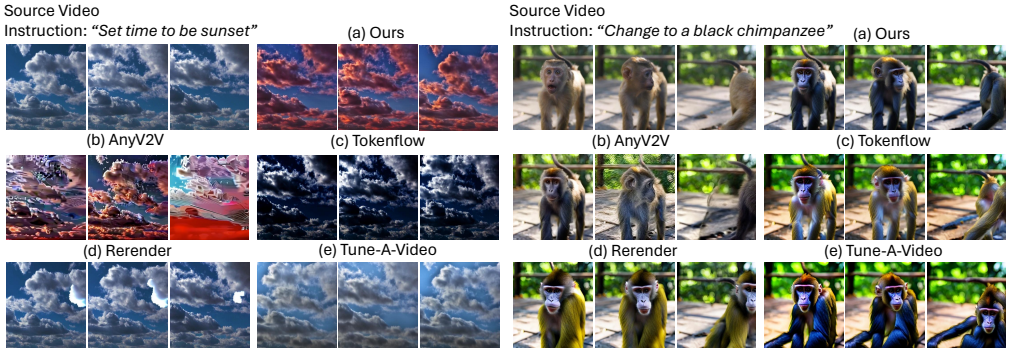


Figure 9: **Qualitative comparison with baselines.** The left side video is a fast-moving cloud, while the right side video is a monkey moving out of the camera. Our model is able to produce consistent editing results.

Qualitative Comparison. In Figure 9, we present two example of video editing. In the first example, the video features a rapidly moving cloud against a blue sky, with the editing directive set to "Set the time to sunset." This task challenges the editing model to deduce the necessary visual alterations. Despite the cloud's swift movement—which places high demands on the model's consistency—our model demonstrates excellent consistency across various frames. Moreover, the Editing Adaptation process enables VIA to effectively align the visual effects with the concept of "sunset." Conversely, other models failed to execute this command adequately. Notably, the AnyV2V model managed to partially achieve the desired visual effect by utilizing the initial frame from VIA. On the right, we present an example of object swapping, where a monkey moves from inside the frame to outside the frame. The challenge lies in ensuring a smooth transition from the whole subject to a partial subject. While other methods often introduce artifacts and exhibit significant inconsistencies between the edited frame and the source video, VIA achieves seamless subject identity swapping, maintaining visual coherence and continuity throughout the transition.

5.3 Human Evaluation

Due to the free form nature of the editing process and the lack of ground truth, we use human evaluation as the quantitative results. Here we sampled 400 videos for human evaluation. We conducted a human evaluation study to assess the performance of our VIA (Ours) against open sourced state-of-the-art baselines, including Rerender, TokenFlow, AnyV2V, Video-P2P, and Tune-A-Video. The evaluation was carried out in three key criteria: Instruction Following, Consistency,

Table 1: **Human evaluation results.** We compare our model with five previous open-source methods from three aspects. 'Tie' indicates the two models are on par with each other.

	Ours	Rerender	Tie	Ours	TokenFlow	Tie	Ours	AnyV2V	Tie	Ours	Video-P2P	Tie	Ours	Tune-A-Video	Tie
Instruction Following	**50.50**	34.00	15.5	**75.75**	16.00	8.25	**56.00**	29.00	15.00	**74.00**	16.25	9.75	**70.25**	20.75	9.00
Consistency	**47.25**	35.00	17.75	**38.00**	31.50	30.5	**53.50**	23.25	23.25	**80.50**	9.50	10.00	**68.75**	20.75	10.5
Overall Quality	**53.50**	29.00	17.5	**61.75**	22.75	15.5	**63.50**	30.00	6.5	**63.75**	22.75	13.5	**56.00**	22.25	21.75

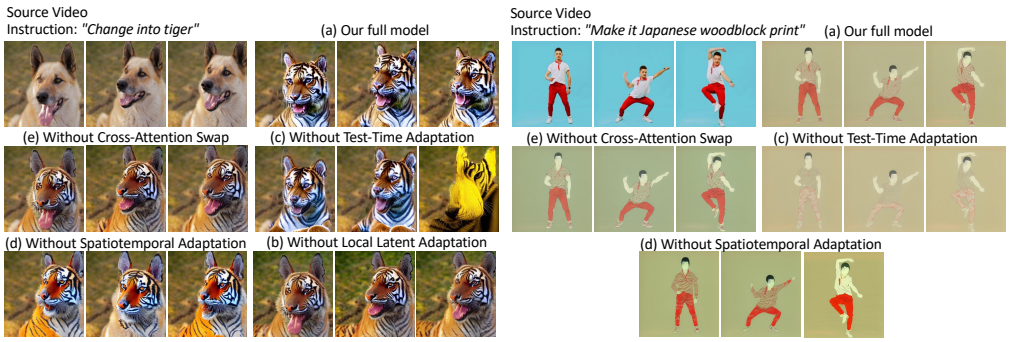


Figure 10: **Ablation Study on components in VIA**. On the left, we present an example of local editing where only the pixels of the dog are altered. On the right, we demonstrate global editing. Without the Local Latent Adaptation process, the background is inevitably affected during editing. Test-time adaptation ensures robust visual effects that accurately adhere to the given instructions. Without the gather-swap technique, object consistency across different frames is compromised. Furthermore, incorporating cross-attention, in addition to self-attention, enhances consistency and reduces artifacts.

and Overall Quality. The results are presented in Table 1. These results highlight the strengths of our proposed method, particularly in Instruction Following and Consistency, while also indicating potential areas for further improvement in terms of Overall Quality compared to certain baselines.

5.4 Ablation Study

In Fig. 10, we demonstrate the impact of various components of VIA on a 20-second video, in which a dog rapidly moves head and shakes body. The editing instruction provided was "Change into a tiger." Our Local Latent Adaptation process effectively identifies the target area and performs precise editing. Additionally, our experiments reveal that the initial edited frames largely determine the overall visual quality, as information from these root frames propagates through the entire video sequence. Test-time adaptation helps the editing model adhere closely to the editing instructions. In the absence of the gather-swap technique and relying solely on cross-frame attention, inconsistencies appear across the frames. Moreover, while self-attention is a standard practice for ensuring frame consistency, we discovered that cross-attention significantly enhances video editing quality. For instance, excluding cross-attention results in less facial alignment with source video.

6 Limitation

While VIA demonstrated impressive performance in video editing, it is not without its limitations. First, it inherits the constraints of the underlying image editing model, which restricts the range of possible editing tasks to those predefined in the image model. Second, the self-tuning process currently necessitates manual selection of a root pair by a human expert. In future iterations, we aim to implement an evaluation model that can automate this selection process.

7 Conclusion

This paper introduced a novel video editing framework that addresses the significant challenges of achieving temporal consistency and precise local edits. Our approach overcomes the limitations of current frame-by-frame methods, ensuring coherent and immersive video experiences. Extensive experiments demonstrate that our framework surpasses existing baselines in terms of temporal dynamics, local edit precision, and overall video aesthetic quality. This advancement opens new possibilities for media production and creative content generation, setting a new standard for future developments in video editing technology.

References

[1] Michael Frierson. *Film and Video Editing Theory*. Routledge, 2018.

[2] Ken Dancyger. *The technique of film and video editing: history, theory, and practice*. Routledge, 2018.

[3] Tao Mei, Xian-Sheng Hua, Linjun Yang, and Shipeng Li. Videosense: towards effective online video advertising. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 1075–1084, 2007.

[4] Nur Kholisoh, Dicky Andika, and Suhendra Suhendra. Short film advertising creative strategy in post-modern era within software video editing. *Bricolage: Jurnal Magister Ilmu Komunikasi*, 7(1):041–058, 2021.

[5] Brendan Calandra, Rachel Gurvitch, and Jacalyn Lund. An exploratory study of digital video editing as a tool for teacher preparation. *Journal of Technology and Teacher Education*, 16(2):137–153, 2008.

[6] Brendan Calandra, Laurie Brantley-Dias, John K Lee, and Dana L Fox. Using video editing to cultivate novice teachers' practice. *Journal of research on technology in education*, 42(1):73–94, 2009.

[7] Wallace Jackson. *Digital video editing fundamentals*. Springer, 2016.

[8] Patrick Schmitz, Peter Shafton, Ryan Shaw, Samantha Tripodi, Brian Williams, and Jeannie Yang. International remix: video editing for the web. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 797–798, 2006.

[9] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023.

[10] 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*, 2023.

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

[12] Bichen Wu, Ching-Yao Chuang, Xiaoyan Wang, Yichen Jia, Kapil Krishnakumar, Tong Xiao, Feng Liang, Licheng Yu, and Peter Vajda. Fairy: Fast parallelized instruction-guided video-to-video synthesis. *CVPR*, 2024.

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[14] Alexander Quinn 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 *ICML*, pages 16784–16804, 2022.

[15] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[16] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022.

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

[18] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

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

[20] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023.

[21] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2023.

[22] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930, 2023.

[23] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089*, 2023.

[24] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023.

[25] Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. *arXiv preprint arXiv:2305.12328*, 2023.

[26] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

[27] Shuzhou Yang, Chong Mou, Jiwen Yu, Yuhan Wang, Xiandong Meng, and Jian Zhang. Neural video fields editing. *arXiv preprint arXiv:2312.08882*, 2023.

[28] Geon Yeong Park, Hyeonho Jeong, Sang Wan Lee, and Jong Chul Ye. Spectral motion alignment for video motion transfer using diffusion models. *arXiv preprint arXiv:2403.15249*, 2024.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015.

[30] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023.

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*, 2023.

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

[33] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023.

[34] Jing Gu, Yilin Wang, Nanxuan Zhao, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, and Xin Eric Wang. Swapanything: Enabling arbitrary object swapping in personalized visual editing. *arXiv preprint arXiv:2404.05717*, 2024.

[35] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *ICLR*, 2024.

[36] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, 2023.

[38] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.

[39] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024.

[40] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022.

[41] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023.

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, volume 36, 2023.

[43] Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, and Xin Eric Wang. Photoswap: Personalized subject swapping in images, 2023.

A Additional Implementation Details

The evaluation was conducted using a collection of online resources and video clips from Panda-70M [39]. VIA can be applied to general image editing frameworks [40, 41, 35]. In this work, we used MGIE [35] as the base image editing model. We set the diffusion step T to 10 and performed spatiotemporal adaptation through all cross-attention and self-attention layers. Our experiments showed that adaptation achieves the best performance when conducted on at least the first 8 steps.

We also observed that increasing the total diffusion step T improves image detail but simultaneously raises the probability of artifacts. Through experimentation, we found that using a value between 5 and 10 generally yields good editing results while maintaining high processing speed. This balance ensures high-quality edits without introducing undesirable visual inconsistencies. For spatiotemporal adaptation, we collect attention variables from four frames.

Test-time Editing Adaptation is a process for refining the editing direction of the underlying model without relying on external data. The pipeline begins with an Edit & Augment step, where a single frame is edited, and transformations are applied to both the source and edited frames to create a training set. Using this dataset, the underlying editing model is fine-tuned to adjust and improve the editing direction. We introduce the following transformations for each image pair, aimed at increasing variability while maintaining the structural integrity of the images: (i) slight rotation (up to ±5 degrees); (ii) translation (up to 5% both horizontally and vertically); and (iii) after applying these transformations, cropping the images to between 75% and 100% of their original size to simulate changes in video sequence framing. Additionally, we apply shearing transformations of up to 10 degrees. These affine transformations introduce realistic variations, simulating the diversity of viewing angles typically encountered across different frames in a video. This approach helps the model adapt to the natural changes in perspective that occur during video sequences. For the tuning process, the training parameter for MGIE is the same as the tuning process of the underlying model. Specifically, we are using a learning rate of 5e-4 with AdamW optimizer, with a batch size of 16 and a total training of 200 steps. Our test-time adaptation process tunes the underlying image editing model towards a fixed editing direction. However, to the best of our knowledge, most video editing methods including the baselines used in this paper use an image generation or video generation model [10, 9, 11]. One exception is one of our baselines, Fairy [12], which uses an image editing model for video editing. However, since it did not open-source the code, it is hard to test the performance of test-time adaptation on other models.

Baseline Implementation primarily follows the publicly available source code. For AnyV2V [26], as it requires an edited first frame, we provide it with the first frame edited by VIA. It inverts the source video into latent space and reconstructs the edited video using the edited frame as a condition. Rerender [38] edits the first frame using a diffusion model, modifies key frames, and interpolates the remaining frames based on the neighboring key frames. TokenFlow [19] inverts each video frame using DDIM to extract tokens and computes inter-frame correspondences via nearest-neighbor search. Keyframes are jointly edited at each denoising step to produce tokens, which are propagated across frames using pre-computed correspondences. The network replaces generated tokens with the propagated ones, iteratively refining the video into the final edited version. Video-P2P [30] employs a diffusion model with a shared unconditional embedding optimized for the reconstruction branch, while the initialized unconditional embedding is used for the editable branch, incorporating the editing instruction. Their combined attention maps generate the target video. Tune-A-Video [10] uses a text-video pair as input and leverages pretrained T2I diffusion models for T2V generation. During fine-tuning, it updates the projection matrices in attention blocks with the standard diffusion training loss. At inference, it generates a new video by sampling latent noise inverted from the input video, guided by a modified prompt. For all methods requiring a new prompt rather than editing instructions, we use ChatGPT to rewrite the prompt. For Fairy [12], as the code is not publicly available, we directly retrieved the video from their official website. For detailed configurations, please refer to their respective papers and open-source code.

From a high level, the difference between VIA and other methods lies in three aspects:

(i) Other models do not consider the local editing process, meaning the editing may fail to faithfully follow the instruction across the entire frame. These methods typically rely on some attention-sharing mechanism without addressing the nuances of video editing.

(ii) For the information-sharing process across different frames, other approaches often directly share information without refinement, whereas VIA employs *gather-and-swap* to **emphasize consistency** in the shared information.

(iii) Their methods are often unsuitable for long videos due to limitations in the backbone architecture. In contrast, our global adaptation process **bypasses these limitations** in current models and hardware (e.g., GPU memory), enabling the editing of videos with up to a few thousand frames.

B Speed Analysis

VIA not only achieves great performance, but also offers impressive speed. The fine-tuning process takes approximately 1 minute, regardless of the video's length. For the global adaptation process, it takes InstructPix2Pix [41] about 1 second per frame, and MGIE [35] around 3 seconds per frame.

Distribution Across GPUs: Once we gather the frames, the editing for all frames can be performed on different GPUs simultaneously, as the frame editing process only depends on the fixed group frames. We utilize 8 GPUs for processing, which helps manage the load effectively.

Total Processing Time for a 600-frame video:

- **MGIE:** 60 (fine-tuning) + $\frac{3\times600}{8} = 285$ seconds.
- **InstructPix2Pix:** 60 (fine-tuning) + $\frac{1\times600}{8} = 135$ seconds.

For the comparison with baselines, where only spatiotemporal adaptation is used (without fine-tuning or local adaptation), the time is:

- **MGIE (without fine-tuning):** $\frac{3\times600}{8} = 225$ seconds.
- **InstructPix2Pix (without fine-tuning):** $\frac{1\times600}{8} = 75$ seconds.

C More Ablation Study

In the main paper, we presented an ablation study on long videos. Here, we demonstrate the impact of various components of VIA on videos less than 20 seconds in duration, where a dog rapidly moves its head and shakes its body. The provided editing instruction was "Change into a tiger." Our Local Latent Adaptation process effectively identifies the target area and performs precise edits. Our experiments also reveal that the initial edited frames largely determine the overall visual quality, as information from these root frames propagates throughout the entire video sequence. Test-time adaptation further ensures that the model adheres closely to the editing instructions.

In the absence of the *gather-and-swap* process, relying solely on cross-frame attention results in inconsistencies across frames. Furthermore, while self-attention is commonly used to maintain frame consistency, we found that cross-attention significantly improves the quality of video editing. For example, when cross-attention is excluded, facial alignment with the source video is reduced, leading to less accurate transformations. In the right part of the experiment, we applied a style change to the video, transforming it into the aesthetic of a Japanese woodblock print. We observed that longer videos exhibit slightly lower visual performance compared to short ones, as minor mismatches can accumulate over a three-minute sequence with approximately 5,000 frames.

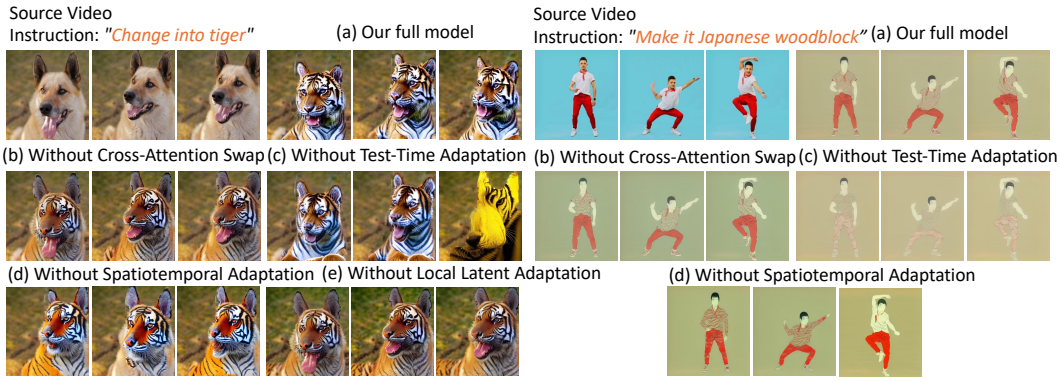


Figure 11: **Ablation study on videos less than 20 seconds.**

Figure 12: **Failure cases.** In the left example, a misalignment occurs during the interaction between the robot and the rock, despite accurately capturing the dance sequence. In the right example, while the driver is seamlessly integrated into the fog, the sequence fails to depict driving out process, leaving the edit incomplete.

D Analysis on Failure Cases

We highlight several failure cases where VIA did not achieve the expected performance, as shown in Fig. 12. The first challenge involves handling complex interactions. In the example on the left, while we successfully captured the intricate body dynamics during a sophisticated dance sequence, a misalignment occurred when the robot was supposed to interact with a rock, leading to inaccuracies at the point of contact. The second challenge relates to temporal dynamics. Although we seamlessly integrated the driver into the fog, the sequence did not show the car emerging from the fog, leaving the scene incomplete. In the future, we plan to incorporate more explicit temporal information into the editing process to better address these issues.

E Automatic Mask Generation

We present an automated mask generation pipeline aimed at enhancing user experience and streamlining the editing process, particularly for large-scale edits. Editing instructions often specify modifications to specific regions, but current end-to-end models tend to alter unintended areas. To address this, we designed an automated pipeline for mask generation, as illustrated in Fig. 13.

First, a Large Vision-Language Model (GPT-4V in our experiment) is prompted to generate a textual description, P, of the region to be modified for each frame. Using this description, we apply the Segment Anything model [32] to extract a mask that accurately delineates the target area for editing. It is important to note that we did not use GPT-4V during comparisons with baselines in the original paper.

In the optimal setting, VIA involves further tuning in the local adaptation process, which some baselines do not utilize. For fairness in comparisons, we degraded our model to use only Spatiotemporal Adaptation during all evaluations. This ensures that our results are directly comparable to baseline models without additional enhancements from local adaptation or the automated mask generation process.

F Performance on Other Backbone

VIA can be equipped with various backbones. Here, we present the performance of another backbone, InstructPix2Pix [41]. As shown in Tab. 2, our model consistently outperforms baselines across multiple metrics. Compared to the MGIE backbone, VIA demonstrates improved *Consistency* performance but slightly lower *Instruction Following* performance. This aligns with the fact that MGIE incorporates an external instruction understanding module [42], which enhances its ability to handle complex editing instructions but diminishes the effect of shared group attention. A similar trend is observed in Tab. 3, where VIA achieves higher performance on *Tem-Con* and *Pixel-MSE* metrics but slightly lower performance on *Frame-Acc*. Furthermore, VIA offers faster editing, as it bypasses the need for the additional instruction understanding process required by MGIE. Here for InstructPix2Pix, we used the same parameter setting as MGIE. In Fig. 14, we present the results on both long and short videos.

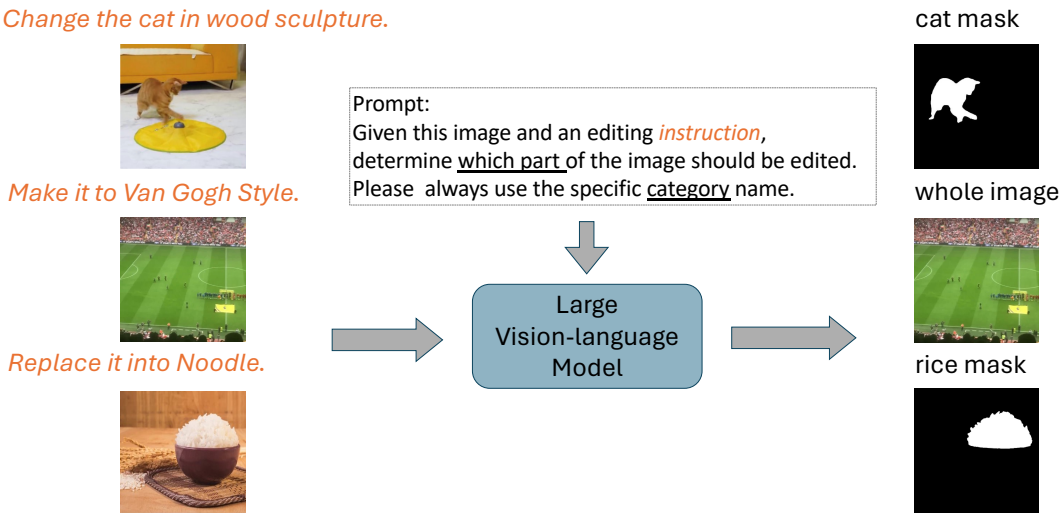


Figure 13: **Automatic mask generation.** A single frame from the video, along with a tailored text prompt encapsulating the editing instruction, is fed into a Large Vision-Language Model (LVLM), such as GPT-4, to generate a text description that specifies the region to be edited. If the designated editing area does not cover the entire image, this text description is then passed into a segmentation model, such as the Segment Anything model, to create a mask for the targeted region. This automated process allows for precise identification of the area to be modified, ensuring that only the relevant portion of the image is edited, while preserving the integrity of the rest of the frame.

Table 2: **Human evaluation results.** We compare our model with five previous open-source methods from three aspects. 'Tie' indicates the two models are on par with each other. Only spatiotemporal adaptation is used when compared with baseline models. Here we used InstructPix2Pix as the backbone.

	Ours	Rerender	Tie	Ours	TokenFlow	Tie	Ours	AnyV2V	Tie	Ours	Video-P2P	Tie	Ours	Tune-A-Video	Tie
Instruction Following	**48.00**	35.00	17.00	**74.00**	18.25	7.75	**53.00**	29.25	17.75	**68.00**	20.25	11.75	**67.00**	22.50	10.50
Consistency	**48.00**	35.50	16.50	**40.00**	31.50	28.50	**54.50**	22.75	22.75	**78.50**	9.50	12.00	**67.75**	19.75	12.50
Overall Quality	**51.00**	28.00	21.00	**59.75**	23.25	17.00	**61.75**	31.50	6.75	**60.25**	24.25	15.50	**51.50**	24.50	24.00

G Comparison on Attention Swapping Process

Attention variables within the U-net of diffusion models have proven to be highly correlated with the generated visual content and are widely used in various editing tasks [40, 33, 43, 30**?**]. In video editing, some methods train models to reconstruct the original videos and swap key attention features during the editing process [26, 30]. Others suggest collecting attention variables independently from individual frame edits and applying them across frames [**?** 12]; however, these independently generated attention variables often lack internal consistency.

In contrast, our recursive *gather* process ensures consistency within the attention group, which is especially crucial for long video generation, where maintaining coherence across thousands of frames is essential. Moreover, unlike previous methods that predominantly rely on self-attention, we also examine the significance of cross-attention layers, as highlighted in the ablation study.

Following the test-time adaptation process, each frame can be edited independently on separate GPUs during the spatiotemporal adaptation phase, significantly reducing the time required, particularly

Table 3: **Automatic evaluation results.** VIA outperforms open-sourced video editing models in automatic metrics. Only spatiotemporal adaptation is used when compared with baseline models. Here we used InstructPix2Pix as the backbone.

	VIA	Rerender	TokenFlow	AnyV2V	Video-P2P	Tune-A-Video
Frame-Acc ↑	**0.862**	0.734	0.587	0.533	0.587	0.601
Tem-Con ↑	**0.985**	0.954	0.932	0.856	0.912	0.927
Pixel-MSE ↓	**0.010**	0.016	0.018	0.026	0.020	0.019
Latency(sec) ↓	**13**	406	450	570	612	529

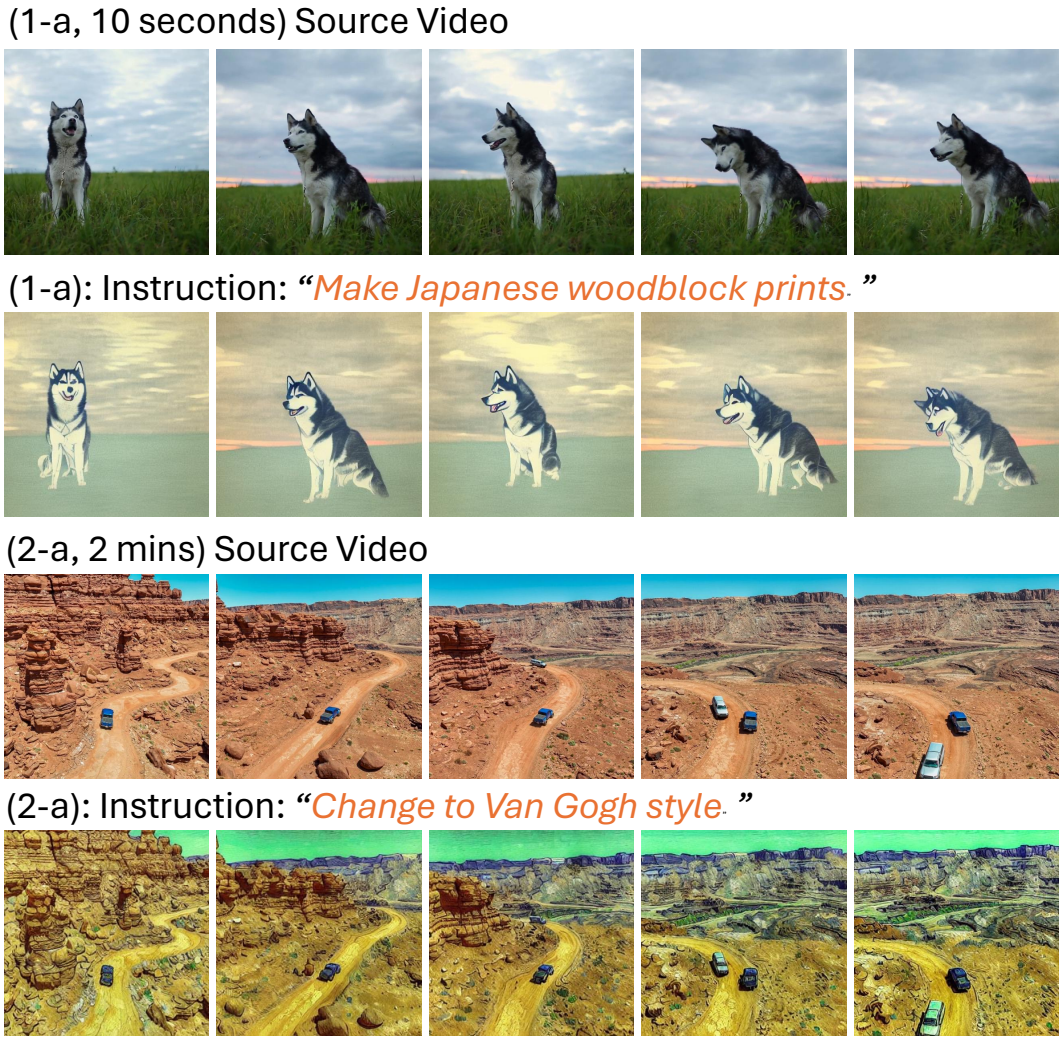


Figure 14: Editing results with InstructPix2Pix. The first one is a 10-second video, and the second one is a 2-minute video.

for long videos. We found that longer videos with more dynamics and scene changes benefit from a larger group size. In this work, we use a group size of 4 for all videos. The attention variable substitution process is performed throughout the entire denoising process, including the classifier-free guidance phase. The *gather* process is essential to the model's success. As shown in Fig. 15, for the same video, using the same random seed and editing instruction, attention gathering produces much more consistent group frames. Without the gathering process, although each frame in the group still follows the instruction, they exhibit different semantic editing directions. With the gathering process, the group maintains internal consistency, and the attention variables from it provide stable guidance for all video frames in the subsequent editing process.

H Blending Comparison

Our proposed Progressive Boundary Integration method differs significantly from traditional blending techniques by dynamically maintaining boundaries across both spatial and temporal dimensions in video editing. Unlike static methods that often cause artifacts like color bleeding or motion inconsistencies, it integrates inverted latent representations progressively, ensuring precise, localized edits without affecting non-targeted areas. The blending method commonly used in the diffusion process could be described as:

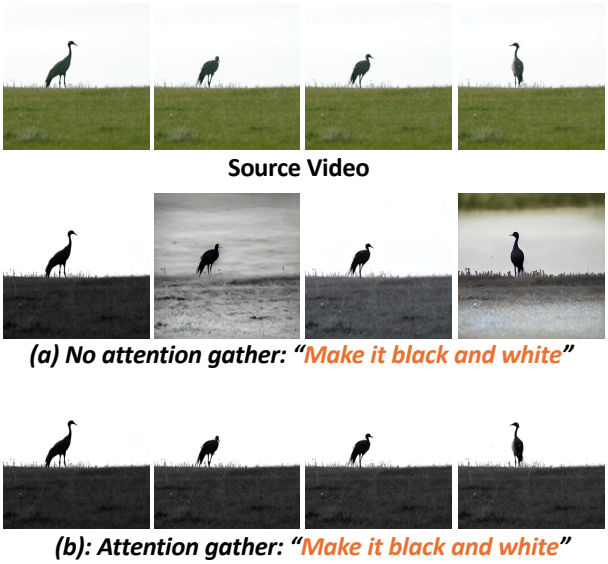


Figure 15: **The edited group frames with&without attention gathering process.** The gathering process ensures in-group consistency, providing a fixed visual editing direction for all frames.

$$\boldsymbol{z}_t^{target} = \mathbf{M}_{\text{src}} \cdot \boldsymbol{z}_t^{edit} + (1 - \mathbf{M}_{\text{src}}) \cdot \boldsymbol{z}_t^{inverted} \tag{10}$$

$$\boldsymbol{z}_{t-1}^{edit} = Sample(\boldsymbol{z}_t^{target}, \Phi, t) \tag{11}$$

While this method works for individual frames, it fails to maintain consistent boundaries for dynamically changing objects in video sequences. This inconsistency leads to variations across frames in the editing area when replacing individual attention with group attention. In contrast, the dynamic mask defined in Equation 6 adjusts adaptively with each time step, allowing the attention to align more effectively with the target area as the diffusion process progresses. In Fig. 16, we present examples of local editing applied to a dog's eyes with the instruction, "Make the eyes glowing." Both Progressive Boundary Integration and direct latent blending successfully preserve the background. However, while the latter performs well on individual frames, it struggles with consistency across the video, as seen in the third frame from the left, where the glowing effect significantly shifts. Experiments demonstrate that our method outperforms standard blending approaches, providing superior control and making it particularly well-suited for video edits that require preserving the integrity of unedited regions.

I Further Improvement with Better Root Frame

In our practice, we observed that a high-quality root frame pair generally leads to improved performance, as illustrated in Fig. 17. In Tab. 4, we show that performance can be further enhanced by incorporating an additional selector. It is important to note that neither a human selector nor an automatic selector was used during the comparison with baselines. By selecting the optimal frame based on editing quality, we ensure that the best possible results are achieved without requiring complex video-level adjustments. This streamlined approach significantly enhances the effectiveness of our method and addresses concerns related to frame selection, allowing for more consistent and visually appealing edits across the video.

J Limitation

While VIA demonstrated impressive performance in video editing, it is not without its limitations. First, it inherits the constraints of the underlying image editing model, which restricts the range of possible editing tasks to those predefined in the image model. Second, the self-tuning process currently necessitates manual selection of a root pair by a human expert. In future iterations, we aim to implement an evaluation model that can automate this selection process.

Figure 16: Comparison between Progressive Boundary Integration and direct latent blending reveals that the former achieves precise and consistent local editing results. For a closer examination, please zoom in on the eye area to observe the editing details.

Table 4: **The selection strategy further improves the results.**

	Manuel	L1	DINO	Random	No Test-time Adaptation
Frame-Acc ↑	**0.891**	0.882	0.887	0.873	0.871
Tem-Con ↑	**0.989**	0.988	0.989	0.983	0.985
Pixel-MSE ↓	**0.0102**	0.0107	0.0108	0.0111	0.0113

K Broader Impact

The advancements introduced by VIA have significant implications across various fields where video editing plays a crucial role. By enabling more precise, consistent, and efficient video editing, particularly for longer videos, VIA opens new possibilities for media production, education, and entertainment, among other domains. Here are some key areas of broader impact:

- **Media and Entertainment**: Our method allows filmmakers, content creators, and advertisers to produce higher-quality, longer-form content more efficiently. This could reduce production time and costs while enhancing the visual appeal and coherence of edited videos. Additionally, artists and creators can experiment with more complex and nuanced edits, fostering greater creative expression.

Figure 17: **Example of frame editing with different seeds.** Edited frames given the source frame on the left and editing instruction "Driving on a river in a forest"

- **Education and Training**: Video is a key tool in educational content, and VIA can significantly improve the quality of instructional videos. Enhanced editing capabilities could lead to better engagement, clearer demonstrations, and more effective communication of ideas. For instance, complex concepts can be explained using tailored visual effects and transitions, making learning more accessible and intuitive.
- **Social Media and User-Generated Content**: As social media platforms increasingly rely on video content, our method can empower non-professional users to create polished, professional-quality videos. This could democratize access to high-end video editing, allowing users without technical expertise to achieve consistent, aesthetically pleasing results.
- **Advertising and Marketing**: In advertising, maintaining brand consistency across video content is critical. VIA's ability to ensure smooth transitions and coherent edits across frames can help marketers maintain the integrity of visual messaging over time, particularly in dynamic, minute-long commercials or social campaigns.
- **AI and Ethical Considerations**: While VIA improves video editing efficiency and quality, it also raises ethical concerns related to video manipulation. The ability to seamlessly edit long videos with high precision could potentially be misused for creating deepfakes or misleading media. As such, there is a need to implement ethical guidelines and detection mechanisms to ensure the responsible use of this technology. Additionally, transparency in editing processes and clear indicators of video manipulation will become increasingly important to prevent misinformation.
- **Environmental Impact**: By improving the efficiency of video editing, VIA reduces the computational resources required for long, complex video edits. This could lead to lower energy consumption, contributing to more environmentally sustainable video production workflows. Reducing the need for re-edits and long processing times could also have positive downstream effects on energy use in large-scale media production.

Overall, the broader impact of VIA extends beyond technical advancements, offering transformative potential across industries while also necessitating careful consideration of ethical and environmental responsibilities.