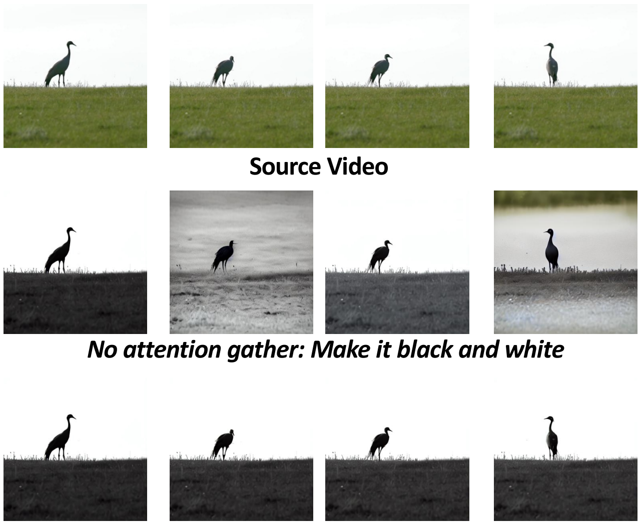

Source Video

No attention gather: Make it black and white

Attention gather: Make it black and white

Figure 8: **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.

A Implementation Details

The evaluation is a collection of online resources and video clip from Panda-70M Chen et al. (2024). VIA could be applied to general image editing framework Hertz et al. (2022); Brooks et al. (2023b); Fu et al. (2024). In this work, we used MGIE Fu et al. (2024) as the image editing model. We set the diffusion step T to be 10, and conduct spatiotemporal adaptation through all cross-attention and self-attention layer. We found the adaptation achieve the best performance when conduct adaptation at least on the first 8 steps. We found that increase the total step T could improve the image details but also increaes the probability of artifacts. We found a value between 1020 usually yeild a good editing results with high speed.

B Architecture Discussion

Local Latent Adaptation . One approach we explored was performing latent $\boldsymbol{z}$ blending only in the final step to merge the two variables. However, this method could introduce artifacts, particularly in the edge areas. Conversely, blending the latent variables without **Progressive Boundary Integration** resulted in images that closely resembled the source frame, thereby minimizing the intended editing effects.

Spatiotemporal Adaptation . After the test-time adaptation process, each frame can be edited on separate GPUs during the spatiotemporal adaptation process, significantly reducing the time required, especially 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. For the attention variable substitution process, we perform it throughout the entire denoising process, including the classifier-free guidance phase. The attention group gathering process is critical to the model's success. As shown in Fig. 8, for the same video, using the same random seed and the same editing instruction, attention gathering yields 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 consistent guidance for all video frames in the later editing process.

	Manuel	L1	DINO	CVS	Random	No Test-time Adaptation
Frame-Acc ↑	**0.891**	0.882	0.887	0.884	0.873	0.871
Tem-Con ↑	**0.989**	0.988	0.989	0.986	0.983	0.985
Pixel-MSE ↓	**0.0102**	0.0107	0.0108	0.0105	0.0111	0.0113

Table 3: **The selection strategy influence on the results.**

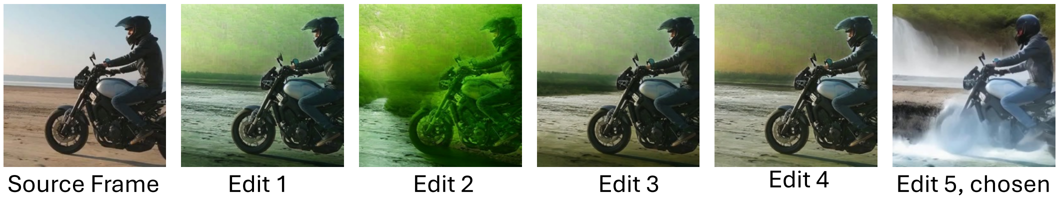

Figure 9: Edited frames given the source frame on the left and editing instruction "Driving on a river in a forest"

C Selection Process

During the frame selection phase, we prioritize the overall editing quality to determine the best frame. In practice, we use 5 different random seeds to generate 5 different frames. Then we select the best frame as the root frame according to the same human evaluation criteria. In human evaluation process, to be a fair comparison, we did not use human in the loop so to have a fair comparison with other models. We demonstrate that this approach indeed enhances the quality of the final output. By selecting the optimal frame based on editing quality, we ensure that the best possible results are achieved without the need for complex video-level adjustments. This streamlined process significantly boosts the effectiveness of our method and addresses the concerns related to frame selection.

D Speed Analysis

VIA not only achieve great performance, but also great speed. For the required mask in local adaptation Approximately 2 seconds when using the GPT4 API to get the editing target. Approximately 0.5 seconds per frame for using Segment Anything. The fine-tuning takes around 1 minute, regardless of the video's length. For the global adaptation process, it takes instructPix2Pix about 1 second per frame, MGIE (with MLLM/LLava) around 3 seconds per frame. Distribution Across GPUs: after we gathered the frames, the editing for all frames could be performed on different GPUs at the same time since the frame editing process only depends on the fixed group frames. We utilize 8 GPUs for processing, which helps in managing the load effectively. Total Processing Time for a 600-Frame Video: MGIE: 2+60+0.5*600/8 + 3*600/8 = 324.5 seconds. InstructPix2Pix: 2+60+0.5*600/8 + 1*600/8 = 174.5 seconds. Note that for the comparison with baselines, where only spatio-temporal adaptation is used (without fine-tuning, local adaptation, or mask preparation), the time is: MGIE: 3*600/8 = 225 seconds. InstructPix2Pix: 1*600/8 =75 seconds. It is worth mentioning that we also tried the recently released segment-anything V2, it works great and the speed for video segmentation is significantly improved (0.02 seconds per frame). Then the segmentation time is negligible. The time calculation above is based on the previous version of Segment-Anything. Lastly, we want to highlight that for all the comparison including human evaluation in the paper, and automatic evaluation, only spatio-temporal adaptation is used, and the local content adaptation including using mask and test-time adaptation are not used in that comparison.

E Usage of External Models

We want to highlight that we did not use GPT4 during comparison with baselines in both the original paper and this rebuttal process. In the optimal setting, VIA involves further tuning and human

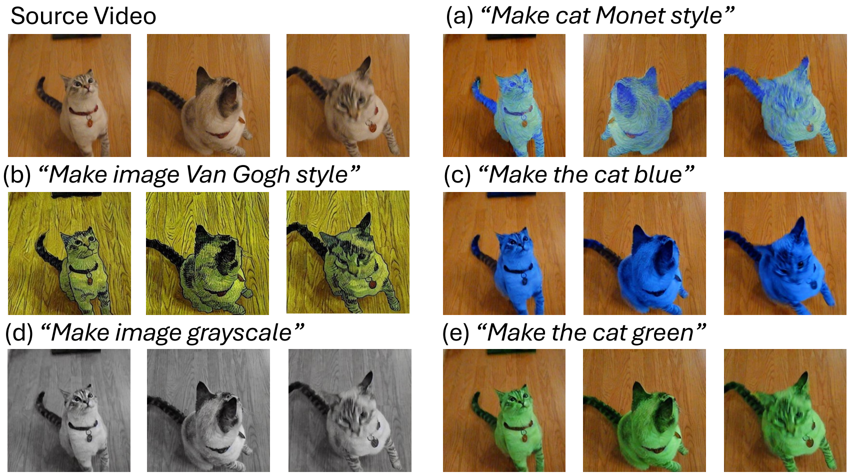
Source Video
(a) *"Make cat Monet style"*
(b) *"Make image Van Gogh style"*
(c) *"Make the cat blue"*
(d) *"Make image grayscale"*
(e) *"Make the cat green"*

Figure 10: **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.

selection in local adaptation process, which are not used in some of the baselines. Therefore, we degrade our model to only use Spatiotemporal Adaptation during all comparisons.

F Analysis on Challenge Cases

While multiple objects with complex interactions are not the focus of our paper, where we directly compare our method with a video presented in their website. We could see that while in the baseline, the dog and the cat could be entangled, our method achieve a much better performance on video editing when there is object intersection.

G Quantitative Analysis

Our analysis, based on 100 videos, highlights the following points: (1) VIA outperforms baselines in terms of both editing quality and latency. Specifically, it ensures smooth transitions in edited videos, even with rapidly moving objects. In contrast, some models, such as AnyV2V, generate noticeable visual artifacts. (2) VIA demonstrates strong performance in adhering to complex instructions. While other models often struggle with complex commands, resulting in degraded performance, our model effectively follows instructions, ensuring that edits are applied consistently across all frames.

H Local Stylization

Fig. 10 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.

I Mask Generation

Editing instructions often specify that only a particular region should be modified, but current end-to-end models frequently alter unintended areas. To solve this, we designed a automated pipeline for mask generation as in Fig. 11. First, a Large Vision-Language Model (LVLM) is prompted to

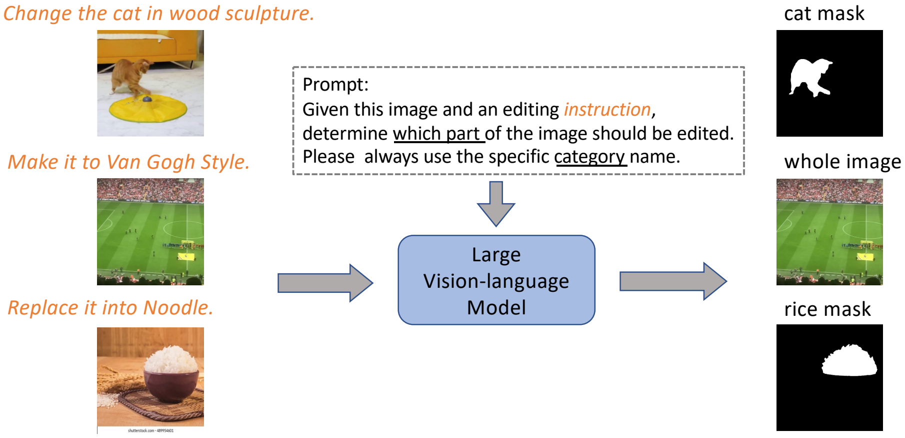
Change the cat in wood sculpture.
Make it to Van Gogh Style.
Replace it into Noodle.
Prompt:
Given this image and an editing instruction, determine which part of the image should be edited. Please always use the specific category name.
Large Vision-language Model
cat mask
whole image
rice mask

Figure 11: **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.

generate a textual description, P, of the region to be modified for each frame. Using this description, P, we apply the Segment Anything model (Kirillov et al., 2023a) to extract a mask that accurately defines the target area for editing.

J 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,

$$\min_\theta \mathbb{E}_{y,\epsilon,t}\left[\left|\left|\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(c_I), c_T)\right|\right|\right] \tag{6}$$

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 (Ronneberger et al., 2015), 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 tokens respectively. Then, the calculated attention map is applied to the value, $\mathbf{V} \in \mathbb{R}^{n \times d}$, describing as follows:

$$\mathbf{Z}' = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V}, \tag{7}$$

$$\mathbf{Q} = \mathbf{Z}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{C}\mathbf{W}_k, \quad \mathbf{V} = \mathbf{C}\mathbf{W}_v, \tag{8}$$

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 (Liu et al., 2023a; Wang et al., 2023; Wu et al., 2024), as shown below:

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

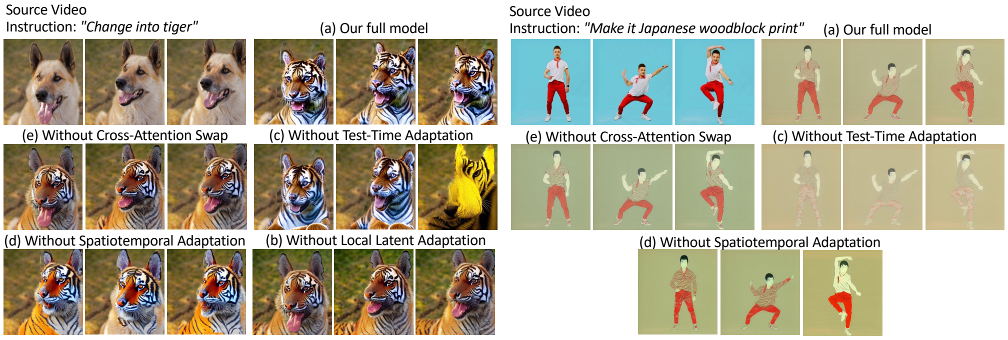
Source Video
Instruction: "Change into tiger"
(a) Our full model
(e) Without Cross-Attention Swap
(c) Without Test-Time Adaptation
(d) Without Spatiotemporal Adaptation
(b) Without Local Latent Adaptation
Source Video
Instruction: "Make it Japanese woodblock print"
(a) Our full model
(e) Without Cross-Attention Swap
(c) Without Test-Time Adaptation
(d) Without Spatiotemporal Adaptation

Figure 12: **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.

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.

K ABLATION STUDY

In Fig. 12, 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.