# Visual Prompting for One-shot Controllable Video Editing without Inversion

Supplementary Material

**Algorithm 1:** Multi-step consistency sampling

**Input:** Consistency model $f_\theta$, textual prompt $p$, noise schedules $\mu_t, \sigma_t$, sequence of timesteps $t_{K:0}$

Sample initial noise $\hat{z}_{t_K} \sim \mathcal{N}(0, I)$

$\hat{z}_0^{(K)} = f_\theta(\hat{z}_{t_K}, t, p)$

**for** $k = K-1 \to 0$ **do**

  Sample $\epsilon \sim \mathcal{N}(0, I)$

  $\hat{z}_{t_k} = \mu_{t_k} \cdot \hat{z}_0^{(k+1)} + \sigma_{t_k} \cdot \epsilon$

  $\hat{z}_0^{(k)} = f_\theta(\hat{z}_{t_k}, t_k, p)$

**Output:** $\hat{z}_0^{(0)}$

**Algorithm 2:** Full procedure of our method for controllable video editing

**Input:** A pre-trained inpainting diffusion model, the first edited frame, latents of source frames $\{z_0^s(i)\}_{i=1}^N$, timesteps $T:0$ and $\mathcal{L}:0$

# CCS process

Sample initial noise $\{\hat{z}_t(i)\}_{i=1}^N \sim \mathcal{N}(0, I)$

Computing initial $\{\hat{z}_0^{(t)}(i)\}_{i=1}^N$ via Eq. (5) in main paper

**for** $t = T-1 \to 0$ **do**

  Updating $\{\hat{z}_0^{(t)}(i)\}_{i=1}^N$ via Eq. (6) in main paper

# TCS process

**for** $\ell = \mathcal{L} \to 0$ **do**

  Updating $\{\hat{z}_0^{(0)}(i)\}_{i=1}^N$ via Eq. (8) in main paper

**Output:** $\{z_0(i)\}_{i=1}^N$

## 1. Preliminary on multi-step consistency sampling of consistency models

**Multi-step consistency sampling** of the consistency models [10, 13] allows generating a sequence of images, where each image corresponds to the model's output at a specific timestep, maintaining content consistency across these images (Alg. (1)). This content consistency arises from the unique mechanism of the multi-step consistency sampling, where the image generated at the current timestep is directly based on the image generated at the previous timestep.

## 2. Our pipeline

The complete procedure of our method is provided in Alg. (2). Besides, as shown in Fig. 1, we also provide a comparison of the pipeline used by our method and previous one-shot controllable video editin methods [6, 9]. Notably, our method eliminates the need for DDIM inversion, which is prone to inaccuracies that can affect the quality of the edited video. By avoiding this error-prone inversion process, our approach achieves more reliable results.

## 3. More implementation details

**Datasets.** We follow Videoshop [6] to use a large-scale video editing dataset derived from MagicBrush dataset [17] to validate the effectiveness of our method. MagicBrush consists of manually annotated image editing data, comprising over 10,000 tuples in the format *(source image, instruction, edit mask, edited image)*. These tuples represent a wide range of edit types, including object addition, replacement, removal, and modifications in action, color, and texture. To adapt MagicBrush into a video dataset, we follow Videoshop to generate videos conditioned on the source images using Stable Video Diffusion [5]. The first frame of each generated video is conditioned to match the corresponding source image, resulting in a video dataset comprising tuples in the format *(source video, instruction, edit mask, edited image)*. In our generated video dataset, each video has a resolution of $1024 \times 576$ and consists of 14 frames. For aesthetic purposes, the frames shown in Fig. 4 of the main paper are at a resolution of $512 \times 512$.

**Baselines.** We compare our method with two state-of-the-art OCVE methods: Videoshop [6] and AnyV2V [9]. Besides, following Videoshop, we also compare our method with five text-based video editing methods: Pix2Video [4], Fatezero [11], Spacetime Diffusion [15], RAVE [7], and BDIA [16]. To create source-target prompt pairs for evaluating text-based video editing methods, we follow Videoshop to caption the first frame of the source video and then modify the described concept to reflect the target (*e.g.*, "a dog at the window" becomes "a bear at the window").

**Evaluation metrics.** To evaluate the quality of the generated edited videos, we follow Videoshop to utilize a comprehensive set of evaluation metrics from five different perspectives. (1) Edit fidelity: We use $\text{CLIP}_{tgt}$ to measure the similarity between the CLIP embeddings of each edited frame and the first edited image. The $\text{CLIP}^+_{tgt}$ score, which utilizes Cotracker [8], is used to measure $\text{CLIP}_{tgt}$ specifically within the edited region. Additionally, we use the TIFA score to assess the semantic alignment between the first edited frame and subsequent edited frames in the video. (2) Source faithfulness: We measure $\text{CLIP}_{src}$ similarity between the source and edited videos, using $\text{CLIP}^+_{src}$ to specif-

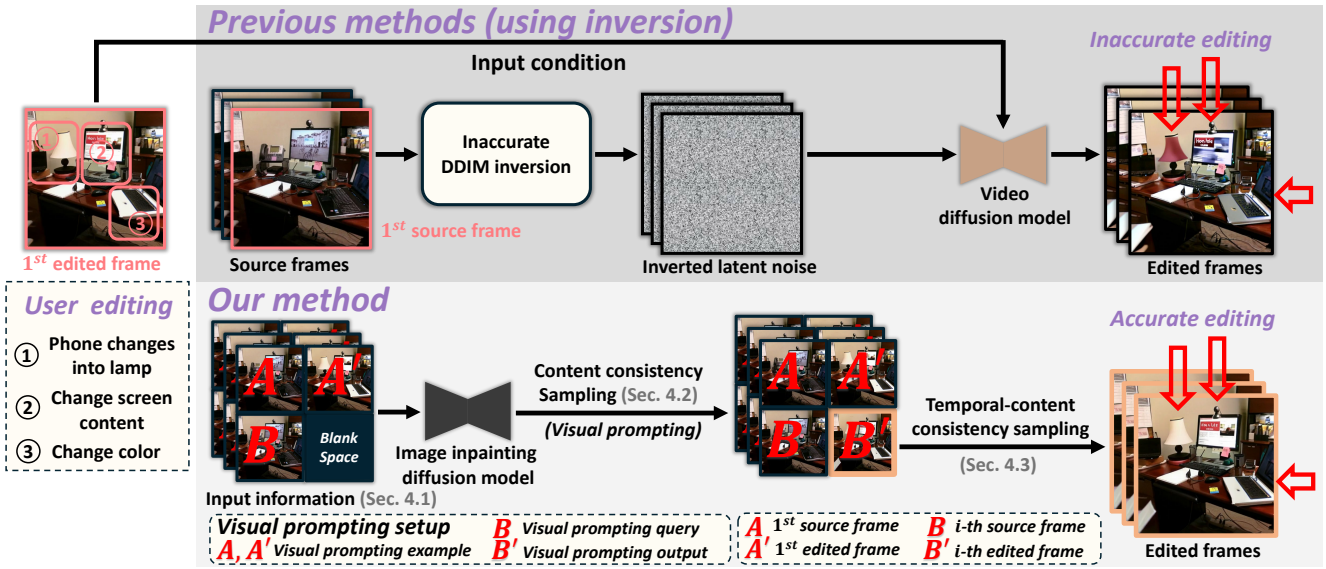


Figure 1. Pervious methods [6, 9] perform one-shot controllable video editing (OCVE) rely on DDIM inversion [12], which can be inaccurate due to the error-prone inversion process [14]. As illustrated in this figure, the edited frames generated by previous methods often fail to maintain the same level of content consistency as the first edited frame, *e.g.*, as shown in the discrepancies in the lamp color, keyboard appearance, and screen content in the edited frames.

Table 1. Results of user study. Best results are in highlight.

| Method | User Preference Rate (%) | |
|---|---|---|
| | Editing quality | Video generation quality |
| AnyV2V [9] | 20.44 | 23.78 |
| Ours | 79.56 | 76.22 |

Table 2. Results of user study. Best results are in highlight.

| Method | User Preference Rate (%) | |
|---|---|---|
| | Editing quality | Video generation quality |
| Videoshop [6] | 27.33 | 32.22 |
| Ours | 72.67 | 67.78 |

ically assess similarity within the unedited regions. Flow is employed to evaluate motion faithfulness by calculating the end-point error (EPE) from optical flow comparisons using RAFT. FLOW$^+$ represents the EPE measured only within the unedited regions. The FVD and SSIM scores are used to assess the overall quality of the generated videos and the quality of individual generated frames, respectively. (3) Temporal consistency: We measure the average CLIP similarity between adjacent frames, referred to as CLIP$_{TC}$. (4) Human evaluation: We ask human evaluators to compare the editing quality of our method with that of the baseline. (5) Efficiency: We measure the average time taken by each video editing method to process a single video.

## 4. More experimental results

**Human evaluation.** We conduct a user study to demonstrate the effectiveness of our method for achieving controllable video editing. A total of 30 individuals from the general public are invited to compare our approach with state-of-the-art OCVE methods, specifically AnyV2V [9] and Videoshop [6]. Each participant is presented with 15 edited videos, which included a diverse range of edits such as object addition, replacement, removal, and modifications in action, color, and texture. Participants are asked to respond to two questions: (1) Which edited video demonstrates better editing quality? (2) Which edited video exhibits superior video generation quality? As illustrated in Tabs. 1 and 2 participants overwhelmingly favored our method over the baseline approaches [6, 9].

**Inpainting diffusion model backbones.** To validate the generalizability of our method, we replace the inpainting diffusion backbone (Stable Inpainting Diffusion 1.5 [1]) with Stable Inpainting Diffusion 2.0 [2]. As shown in Tab. 3, we evaluate the effect of different backbones on our method by conducting comparisons on the generated dataset. The results indicate that the editing performance is largely unaffected by this change in backbone. Given the efficiency considerations, we use Stable Inpainting Diffusion 1.5 [1] as the backbone for our approach.

## 5. Video sample

Our additional video samples are available on the project page for further visual evaluation.

Table 3. Quantitative comparisons of our method with different backbones. Best results are highlighted. "+" indicates that the metric is used to evaluate the edited region, whereas "−" indicates that the metric is used to evaluate the unedited region. Following Videoshop [6], we use Cotracker [8] to identify the edited and unedited regions. (T.C. = Temporal Consistency; E. = Efficiency)

| Method | Edit Fidelity | | | Source Faithfulness | | | | | | T.C. | E. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\text{CLIP}_{\text{tar}}\uparrow$ $(\times 10^{-2})$ | $\text{CLIP}^{+}_{\text{tar}}\uparrow$ $(\times 10^{-2})$ | TIFA $\uparrow$ $(\times 10^{-2})$ | $\text{CLIP}_{\text{src}}\uparrow$ $(\times 10^{-2})$ | $\text{CLIP}^{+}_{\text{src}}\uparrow$ $(\times 10^{-2})$ | Flow $\downarrow$ $(\times 10^{-1})$ | $\text{Flow}^{-}\downarrow$ $(\times 10^{-1})$ | FVD $\downarrow$ $(\times 10^{2})$ | SSIM $\uparrow$ $(\times 10^{-2})$ | $\text{CLIP}_{\text{TC}}\uparrow$ $(\times 10^{-2})$ | time (s) $\downarrow$ $(\times 10^{0})$ |
| Ours w/ Stable Diffusion Inpainting 2.0 | 90.3 | 87.9 | 69.2 | 93.4 | 96.5 | 22.3 | 9.4 | 15.2 | 69.1 | 97.0 | 21 |
| Ours w/ Stable Diffusion Inpainting 1.5 | 90.1 | 88.2 | 69.1 | 93.2 | 96.6 | 21.9 | 9.2 | 15.2 | 69.2 | 97.1 | 19 |

## 6. Limitation

From the quantitative results (see Tab. 1 in the main paper), it is clear that our method surpasses previous state-of-the-art OCVE methods [6, 9] across all metrics related to Edit Fidelity and Temporal Consistency. However, our method is still slightly behind on certain metrics reflecting Source Faithfulness when compared to Videoshop [6]. This can be attributed to the fact that Videoshop employs a more powerful video diffusion model with 1.50 billion parameters, whereas we utilize the Stable Diffusion Inpainting 1.5, which contains only 0.86 billion parameters. We anticipate that future developments in more advanced open-source inpainting diffusion models could improve the performance of our approach. Furthermore, since our test dataset is generated using SVD [3], the resulting test videos contain objects with relatively small motion.

## References


[1] Stability AI. Stable diffusion inpainting. https://github.com/Stability-AI/stablediffusion, 2022. 2

[2] Stability AI. Stable diffusion inpainting 2.0. https://huggingface.co/stabilityai/stable-diffusion-2-inpainting, 2022. 2

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[4] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023. 1

[5] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 1

[6] Xiang Fan, Anand Bhattad, and Ranjay Krishna. Videoshop: Localized semantic video editing with noise-extrapolated diffusion inversion. In *European conference on computer vision*. Springer, 2024. 1, 2, 3

[7] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6507–6516, 2024. 1

[8] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 1, 3

[9] 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. 1, 2, 3

[10] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 1

[11] 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. 1

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[13] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 1

[14] Yangyang Xu, Wenqi Shao, Yong Du, Haiming Zhu, Yang Zhou, Ping Luo, and Shengfeng He. Task-oriented diffusion inversion for high-fidelity text-based editing. *arXiv preprint arXiv:2408.13395*, 2024. 2

[15] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8476, 2024. 1

[16] G. Zhang, J. P. Lewis, and W. B. Kleijn. Exact diffusion inversion via bi-directional integration approximation. In *European conference on computer vision*. Springer, 2024. 1

[17] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 1