FramePainter: Endowing Interactive Image Editing with Video Diffusion Priors

Yabo Zhang[1]   Xinpeng Zhou[1]   Yihan Zeng[2]   Hang Xu[2]   Hui Li[1]   Wangmeng Zuo[1, ✉]
[1]Harbin Institute of Technology   [2]Huawei Noah's Ark Lab
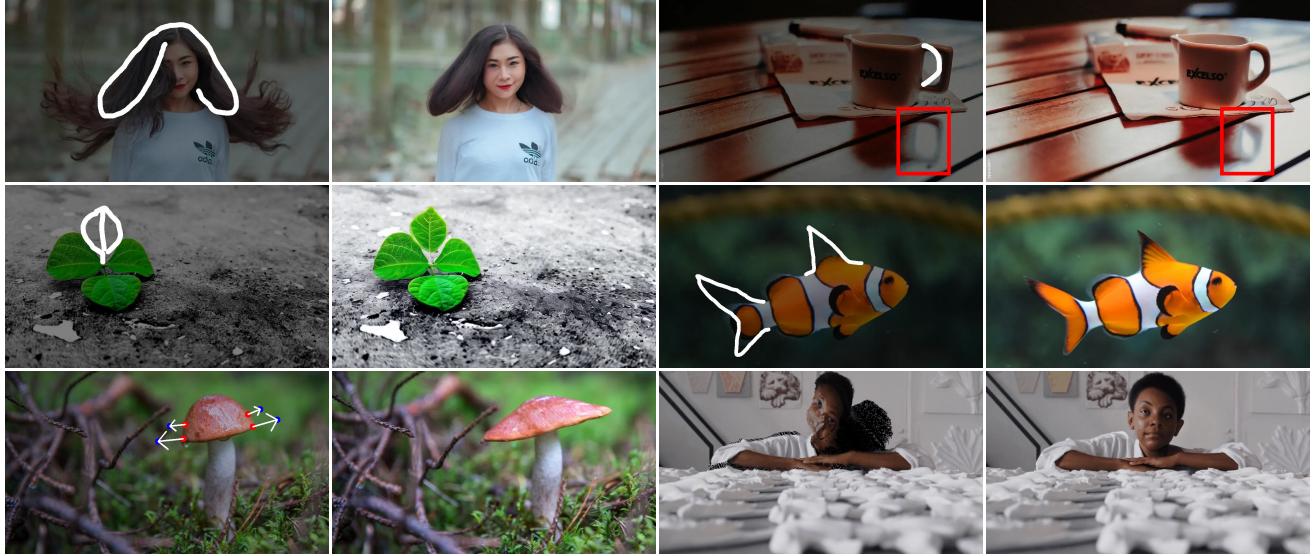
Figure 1. **Examples of FramePainter**. FramePainter allows users to manipulate images through intuitive visual instructions like drawing sketches, clicking points, and dragging regions. Benefiting from powerful video diffusion priors, it not only enables intuitive and plausible edits in common scenarios (*e.g.*, adjust the reflection of the cup in red box), but also exhibits exceptional generalization in out-of-domain cases, *e.g.*, transform the clownfish into shark-like shape.

Abstract

Interactive image editing allows users to modify images through visual interaction operations such as drawing, clicking, and dragging. Existing methods construct such supervision signals from videos, as they capture how objects change with various physical interactions. However, these models are usually built upon text-to-image diffusion models, so necessitate (i) massive training samples and (ii) an additional reference encoder to learn real-world dynamics and visual consistency. In this paper, we reformulate this task as an image-to-video generation problem, so that inherit powerful video diffusion priors to reduce training costs and ensure temporal consistency. Specifically, we introduce FramePainter as an efficient instantiation of this formulation. Initialized with Stable Video Diffusion, it only uses a lightweight sparse control encoder to inject editing signals. Considering the limitations of temporal attention in handling large motion between two frames, we further propose matching attention to enlarge the receptive field while encouraging dense correspondence between edited and source image tokens. We highlight the effectiveness and efficiency of FramePainter across various of editing signals: it domainantly outperforms previous state-of-the-art methods with far less training data, achieving highly seamless and coherent editing of images, e.g., automatically adjust the reflection of the cup. Moreover, FramePainter also exhibits exceptional generalization in scenarios not present in real-world videos, e.g., transform the clownfish into shark-like shape. Our code will be available at https://github.com/YBYBZhang/FramePainter.

1. Introduction

Diffusion models have achieved remarkable success in generating exceptional and photorealistic images from natural language descriptions [45, 47–49]. Compared to generating images from scratch, the users prefer utilizing these models

to edit captured photographs or designed images. While text-guided image editing demonstrates promising potential [5, 7, 17, 22, 32, 35], it is constrained by the ambiguity of language instructions and the lack of precise spatial control, *e.g.*, failing to accurately adjust the shape, position, or posture of a human. In contrast, interactive image editing [1, 11, 25, 41, 51] offers a more flexible and precise solution, which supports more intuitive operations like drawing sketches, clicking points, and dragging regions.

Most existing approaches [1, 37, 50, 51] treat interactive image editing as an image-to-image generation task and leverage pre-trained text-to-image diffusion models [45, 47–49] as foundation models. During finetuning, they [1, 51] usually construct source and target image pairs from real-world videos, which provide sufficient observations of how objects change during physical interactions, *e.g.*, a man raising his head. However, in the absence of priors about real-world dynamics, these methods necessitate enormous training samples to achieve plausible and precise control over images. Additionally, they require additional reference branches (*e.g.*, the Reference U-Net [18] or IP-Adapter [62]) to maintain appearance consistency, further increasing model complexity and training costs.

In this paper, we reformulate interactive image editing as an image-to-video generation problem: the source image acts as the first frame, and an editing signal directs the generation of a video comprising the source and target images. Our novel formulation enables the image editing models to leverage real-world dynamic priors from pre-trained video diffusion models [4], thereby potentially reducing both data and computational requirements compared to existing approaches. Moreover, leveraging the features of the first frame for identity preservation, it eliminates the need for additional reference encoders and simplifies the model architecture.

Within this formulation, we propose *FramePainter* to achieve flexibly and precisely interactive image editing using powerful video diffusion priors. Specifically, FramePainter is initialized with Stable Video Diffusion (SVD) [4] and uses a lightweight sparse control encoder to inject an editing signal (*e.g.*, a sketch image) into the U-Net. Since the temporal attention of SVD struggles with handling large motion between two images, we further introduce *matching attention* to encourage dense correspondence between edited and source image tokens. In particular, matching attention extends spatial attention along the temporal axis to enlarge the receptive field between images, and is implemented as an auxiliary branch to complement spatial attention. To capture fine-grained visual details more precisely, we optimize matching attention by aligning its attention weights with tracking results from CoTracker-v3 [21]. During inference, matching attention does not require tracking results as input, but can accurately query the corresponding source image token for each edited image token (refer to Fig. 8 for visualization).

To verify the effectiveness of FramePainter, we construct thousands of image pairs from high-quality videos and split them into training and evaluation datasets. Concretely, each image pair is randomly sampled and will be saved if their optical flow magnitude is large enough on local area (*i.e.*, significant object movement). Given two appropriate images, we extract corresponding editing signals with their optical flow or trackings, *e.g.*, sketch images and dragging points. Through extensive experiments on a wide range of editing signals, FramePainter not only significantly outperforms training-free methods in editing plausibility and visual consistency, but also surpasses training-based methods with substantially lower training costs. Moreover, FramePainter demonstrates remarkable generalization in scenarios that are absent in real-world videos, such as transforming a clownfish into a shark-like shape.

Our core contributions are summarized as:

- We reformulate interactive image editing as an image-to-video generation task, and introduce FramePainter to facilitate flexible and precise image manipulation using powerful video diffusion priors.
- To capture fine-grained visual appearance more precisely, we propose matching attention to encourage dense correspondence between edited and source image tokens.
- The experiments demonstrate that FramePainter achieves superior performance across various editing signals with far less training costs, while also showcasing exceptional generalization to unseen scenarios in real-world videos.

2. Related Work

Image and Video Diffusion Models. Diffusion models have become the de facto standard in generative modeling and drive significant advancements in both image [2, 39, 45, 47–49, 65] and video generation [3, 4, 8, 12, 16, 23, 28, 30, 55, 60, 61, 63, 67, 68, 71]. In the field image synthesis, powerful foundational models such as Stable Diffusion [45, 48] demonstrate unprecedented capabilities in producing realistic and diverse images and are thus widely applied to downstream tasks like text-guided image editing [6, 7, 33, 34, 43, 54] and controllable generation [19, 26, 29, 38, 57, 58, 66, 69, 70]. Built upon pre-trained image diffusion models, video diffusion models [3, 4, 8, 12, 16, 55, 60, 61, 63, 67, 71] inherit their capability to generate high-quality video frames. By introducing temporal modules and training on large-scale videos, they excel in understanding and recreating real-world dynamic processes by capturing temporal consistency and natural motion patterns. However, due to the limited receptive field, it is challenging for temporal modules to deal with large movements in two frames, which can be mitigated by our proposed matching attention.
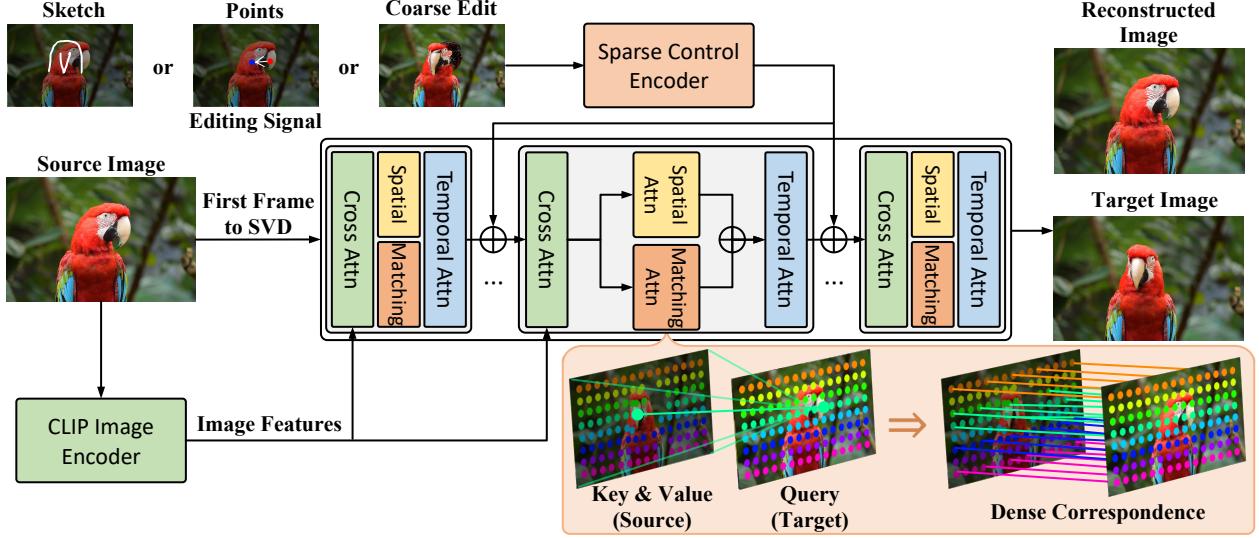
Figure 2. **Overview of FramePainter.** Reformulating image editing as an image-to-video generation task, FramePainter takes a source image and an editing instruction as the first frame and control guidance, and produces a two-frame video comprising of reconstructed and target images. To improve visual consistency of two images involving large motion, matching attention is proposed to enlarge the receptive field and encourage dense correspondence between target and source image tokens.

Interactive Image Editing. The breakthroughs of diffusion models in text-to-image generation have significantly propelled advancements in the domain of image editing. Among them, text-guided image editing [5, 7, 17, 22, 32, 34, 35, 43, 54] are limited by the inherent ambiguity of language instructions and the absence of precise spatial control, while interactive image editing [1, 11, 25, 27, 37, 40, 41, 51] allows users to more flexibly and precisely manipulate images through intuitive instructions, such as drawing sketches, clicking points, and dragging regions [1]. An earlier work DragGAN [41] obtains promising point-based editing with motion supervision and point tracking but is limited, but is constrained by the inherent capacity of StyleGAN. DragDiffusion [50] and DragonDiffusion [37] achieve open-domain point-based editing with the power of text-to-image diffusion models. Magic Fixup [1] proposes a novel way to modify images, where an image is first coarsely edited by users and then transformed to realistic image through the network. Existing works are tailored to a single type of editing signal and primarily focus on clicking points or dragging regions. By contrast, FramePainter is capable of various editing signals and introduces drawing sketch as a more intuitive and convenient editing signal.

Video Diffusion Priors for Generative Tasks. Video diffusion models [3, 4, 8, 12, 16, 55, 61, 63, 67, 71] possess stronger priors than image diffusion models when modeling real-world dynamic processes, particularly in capturing temporal consistency and plausible physical interactions. While most existing works utilize these models for video-

related downstream tasks [14, 15, 20, 31, 42, 53, 59], a few studies have explored the potential of directly learning from videos to enhance image-related tasks. For example, Anydoor [9] diversifies object poses and viewpoints by borrowing knowledge from videos. MagicFixup [1] and Lightning-Drag [51] select appropriate video frames as supervision signals of image editing task. However, since these models are built upon image diffusion models, they requires extensive training data and additional reference encoders to learn video knowledge from scratch. Differently, FramePainter pioneers the integration of video diffusion priors into image editing task, significantly reducing training costs and simplifying the model architecture.

3. Methodology

Interactive image editing provides users intuitive operations to flexibly and precisely manipulate real-world images. In this work, we reformulate it as an image-to-video generation task, and introduce *FramePainter* to inherit real-world dynamic priors from video diffusion models. To improve fine-grained visual consistency, we propose *matching attention* to encourage each edited image token align with its corresponding source image token. To train our model, we construct image pairs along with visual editing instructions from high-quality videos.

3.1. Image Editing as an Image-to-Video Task

Interactive image editing aims to simulate real-world changes based on user-provided editing instructions, such

3

| Source Image | Target Image | Editing Signal |

Figure 3. **Collected samples from videos.** We present three types of editing signals from top to bottom: drawing sketches, click points, and dragging regions.

as sketching, clicking, or dragging. Since video corpora provide abundant observations of how the world changes through physical interactions, they can serve as ideal supervision signals for learning such editing behaviors [1, 51]. However, existing methods [1, 51] are typically initialized with pre-trained image diffusion models and lack real-world dynamic priors, thus require a massive number of training samples and additional reference encoders to achieve plausible edits and visual consistency. In contrast, we reformulate interactive image editing as an image-to-video generation task and introduce *FramePainter* to mitigate above issues.

Fig. 2 illustrates the pipeline of FramePainter. Initialized with SVD [4], FramePainter takes a source image I^{src} (*i.e.*, the first frame to SVD) and an editing signal s as input, and then produces a two-frame video that consists of reconstructed and target images. Inspired by ControlNext [44], we use a lightweight sparse control encoder with multiple ResNet blocks to efficiently encode editing signals. Notably, we only inject editing signals into target image features to avoid affecting the reconstruction of source image. The model ϵ_θ and control encoder ϕ are finetuned with the diffusion loss:

$$\mathcal{L}_{diff} = \mathbb{E}_{z_0, s, t, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(z_t, t, I^{src}, \phi(s)) \|_2^2 \right], \quad (1)$$

where z_0 is concatenated from source image latent z_0^{src} and target image latent z_0^{tgt} along the temporal axis. Following SVD [4], we use two ways to preserve the visual details of I^{src}: (i) inject the features encoded by CLIP image encoder [46] into cross-attention modules, and (ii) concatenate z_0^{src} with per-frame noise latent in channel dimension and then feed them to the denoising U-Net.

3.2. Matching Attention for Dense Correspondence

Since SVD relies on 1-D temporal attention to ensure frame consistency, its small receptive field struggles to preserve the identity of objects involving large movements, particularly when only two frames are available (*i.e.*, in image editing task). Existing works [23, 68] inflate spatial attention into cross-frame attention to enlarge the receptive field, but still fail to achieve fine-grained consistency in appearance [10, 13]. To alleviate it, we propose matching attention to facilitate dense correspondence between target and source image tokens.

As shown in Fig. 2, matching attention is implemented as an auxiliary branch of spatial attention to capture visual details from source image. Given source image tokens $z_t^{src} \in \mathbb{R}^{N \times d}$ and target image tokens $z_t^{tgt} \in \mathbb{R}^{N \times d}$ at timestep t, matching attention only takes z_t^{tgt} as query and z_t^{src} as key and value:

$$\mathbf{A}^{match} = \text{Softmax}(\frac{\mathbf{Q}^{match}(\mathbf{K}^{match})^T}{\sqrt{d}}), \quad (2)$$

$$\mathbf{O}^{match} = \mathbf{A}^{match} \cdot \mathbf{V}^{match}, \quad (3)$$

where $\mathbf{Q}^{match} = \mathbf{W}_Q^{match} \cdot z_t^{tgt}$, $\mathbf{K}_{match} = \mathbf{W}_K^{match} \cdot z_t^{src}$, and $\mathbf{V}^{match} = \mathbf{V}_V^{match} \cdot z_t^{src}$. Next, we pad the output $\mathbf{O}^{match}$ with zeros and add it to the output of spatial attention $\mathbf{O}^{spat}$ as follows:

$$\mathbf{O}^{final} = \mathbf{O}^{spatial} + [\mathbf{0}, \mathbf{O}^{match}], \quad (4)$$

During training, matching attention copies the weights from spatial attention to inherit its implicit knowledge on image correspondences [52]. To encourage more precise correspondence, we employ CoTracker-v3 to extract tracking results, including correspondence matrix $\mathbf{C} \in [0, 1]^{N \times N}$ and visible mask $\mathbf{M} \in [0, 1]^{N \times N}$:

$$\mathbf{C}[i, j] = \begin{cases} 1, & \text{if } z_t^{tgt}[i] \text{ corresponds to } z_t^{src}[j] \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

$$\mathbf{M}[i, j] = \begin{cases} 1, & \text{if } z_t^{tgt}[i] \text{ has visible tokens in } z_t^{src} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Then, we use them optimize the attention weights of matching attention as:

$$\mathcal{L}_{match} = \| \mathbf{M} \cdot (\mathbf{A}^{match} - \mathbf{C}) \|_2^2, \quad (7)$$

Finally, the overall learning objective is defined as:

$$\mathcal{L} = \mathcal{L}_{diff} + \lambda_{match} \cdot \mathcal{L}_{match}. \quad (8)$$

where λ_{match} controls the scale of matching loss and is set to 1.0 by default. As illustrated in Fig. 8, matching attention enables each target token to focus more accurately on its corresponding source token than vanilla cross-frame attention, *i.e.*, translating sparse editing signals into dense correspondences.

4

Table 1. **Quantitative comparisons across different types of visual editing instructions.** Despite using fewer than 10% or 1% training samples than previous state-of-the-art methods [1, 51], FramePainter surpasses alternative approaches across all editing signals. The best results are **bolded**.

Method	Editing Signal	Training Samples	CLIP-FID ($\downarrow$)	LPIPS ($\downarrow$)	SSIM ($\uparrow$)
MasaCtrl + ControlNet	Sketch	None	17.933	0.302	0.655
FramePainter (Ours)	Sketch	20k	**7.783**	**0.140**	**0.859**
Magic Fixup	Coarse Edit	2,500k	8.757	0.166	0.855
FramePainter (Ours)	Coarse Edit	20k	**7.573**	**0.132**	**0.888**
DragDiffusion	Points	None	9.192	0.187	0.811
LightningDrag	Points	220k	9.894	0.214	0.794
FramePainter (Ours)	Points	20k	**8.513**	**0.166**	**0.825**

3.3. Constructing Samples from Video Data

Videos capture a wide range of real-world transformations, including object movements, pose variations, and perspective changes, serving as natural supervision signals for image editing task. Considering the goal of manipulating local regions, we curate high-quality videos with static camera to better support subsequent data construction [51]. Fig. 3 presents the examples of various editing signals.

Sampling Suitable Image Pairs. To ensure sufficient movements, we randomly sample two frames from the video with a time interval greater than one second. With the optical flow predicted by SEA-RAFT [56], we filter out frame pairs whose flow magnitude are excessively small or large, and obtain 22,000 image pairs in total.

Extracting Visual Editing Instructions. We begin to extract optical flow and tracking points from target image to source image with SEA-RAFT [56] and CoTracker-v3 [21], respectively. For each type of editing signal, we design different algorithms to construct them: *(i)* Sketches. We utilize a Sobel filter [64] to directly extract the edges of the optical flow, which are then used as the sketch signals. *(ii)* Dragging points. Following LightningDrag [51], we randomly sample target points with a probability weighted by optical flow magnitude, and then find their corresponding source points according to tracking results. *(iii)* Coarsely edited images. Following Magic Fixup [1], we employ the softmax splatting algorithm to warp the source image, where the warped images are considered as producing a coarsely edited image that aligns with the target.

4. Experiments

4.1. Experimental Setup

Implementation Details. We initialize FramePainter with Stable Video Diffusion v1.1 and finetune it on our collected training samples for each editing signal. Following SVD, the height and width of training images are set to 576 and 1024 respectively. We train the model on two A6000 GPUs

Table 2. **User preference study.** The numbers denote the percentage of raters who favor the images edited by FramePainter over other baselines.

Method Comparison	Visual Cons.	Edit Acc.	Image Qual.
Ours vs. MasaCtrl	88.3%	82.0%	83.2%
Ours vs. Magic Fixup	71.1%	72.4%	76.5%
Ours vs. LightningDrag	73.2%	68.9%	72.7%

with a total batch size of 4. The model is optimized for 20,000 iterations with a learning rate of 1×10^{-5} using the AdamW algorithm. During inference, we adopt euler discrete sampling with 25 steps by default.

Training and Evaluation Datasets. We partition the collected samples into training and evaluation datasets, consisting of $\sim 20,000$ and 200 samples, respectively. When training the model for each type of editing signal, the only difference is the editing signal used, with the image pairs remaining exactly the same. Compared to existing evaluation benchmarks [40], our evaluation benchmark not only includes source images and editing signals but also provides target images, enabling a more comprehensive assessment of the edited results.

Evaluation Metrics. We use three metrics to evaluate the performance of FramePainter: *(i)* CLIP-FID [24] evaluates the semantic alignment between edited images and target images using CLIP embeddings. Lower scores indicate better semantic consistency. *(ii)* LPIPS measures perceptual similarity between images based on deep feature representations. Lower values reflect closer visual similarity. *(iii)* SSIM assesses structural similarity by comparing luminance, contrast, and structure. Higher scores indicate better preservation of structural details.

4.2. Comparisons with Baselines

Qualitative Results. Fig. 4 presents the visual comparisons of edited images under different editing signals. In the top of Fig. 4, FramePainter not only maintains the vi-
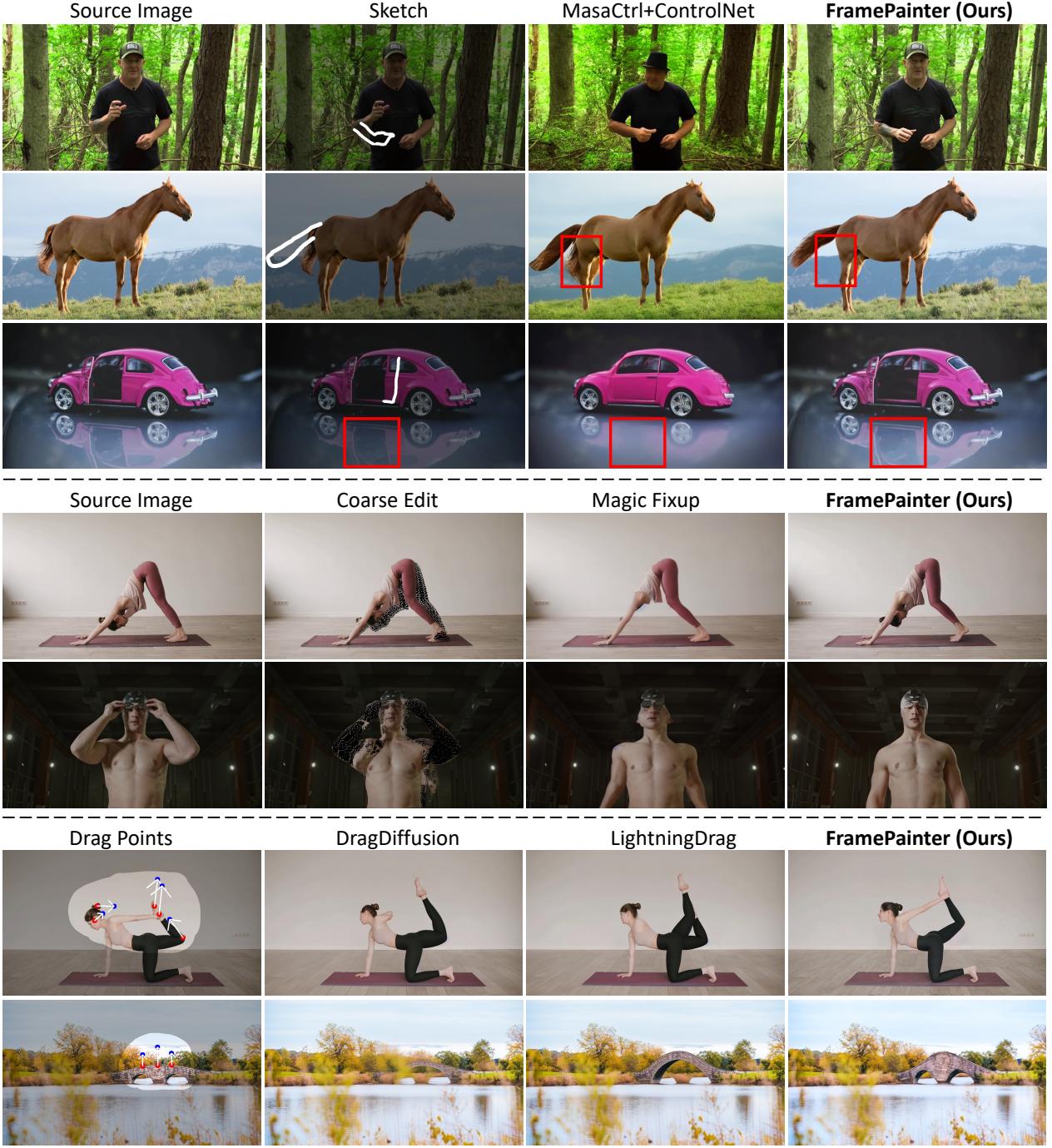
Figure 4. **Qualitative comparisons across different visual editing instructions.** Compared to the baselines, FramePainter not only achieves more coherent and plausible editing results, but also automatically polishes the edited images to meet real-world dynamics, *e.g.*, remove duplicate tail and adjust car door in mirror (highlighted in red box). We note that LightningDrag and DragDiffusion require users to provide additional masks, whereas FramePainter does not.

sual appearance (*e.g.*, the color of hat), but also achieves more natural and plausible edits than MasaCtrl+ControlNet. With rich priors from pre-trained video diffusion models, the removal of the car door automatically updates its mir-

ror reflection (highlighted in red box), ensuring the visual consistency and realism. In the middle of Fig. 4, the images refined by FramePainter contain more realistic details, while the results from MagicFixup exhibit noticeable arti-

6

Figure 5. **Emerging capabilities of FramePainter.** Although FramePainter is trained on image pairs from real-world videos, it demonstrates several emerging capabilities as a convenient tool: **(i)** Supporting highly intuitive and simplified instructions. **(ii)** Offering precise control over complex editing signals. **(iii)** Generalizing well to out-of-domain cases, such as shape transformation.

facts, *e.g.*, missing head and distorted face. As shown in the bottom of Fig. 4, FramePainter excels at maintaining the structural integrity of objects than the alternative baselines. For example, the baselines produce edited images missing arms or bridge arches, as they focus on keeping appearance with LoRAs or ReferenceNet but overlook structural consistency.

Quantitative Results. Table 1 compares the quantitative results in each type of editing signal. As one can observe, FramePainter significantly outperforms the state-of-the-art approaches under all editing settings, which is consistent with the qualitative results. In row $1 - 2$ of Table 1, it surpasses the performance of MasaCtrl+ControlNet in a large margin. From row 3 to 7, despite using only 1% or even 0.1% of the training data required by previous methods, FramePainter achieves superior performance, highlighting the effectiveness and efficiency of our novel formulation.

User Study. To verify the effectiveness of FramePainter, we conduct a user study in 100 samples consists of source images and editing signals. We randomly produce 5 edited images for each sample and obtain 500 images in total. For each type of editing signal, we provide the rater a source image, an editing signal, and two edited images from different methods (in random order). Then, they are asked to select the better edited image for each of three perspectives: (i) visual consistency, (ii) edit accuracy, and (iii) image quality. Table 2 summarizes the voting results of raters. As one can see, the raters strongly favor the images edited by FramePainter rather than the other baselines from all three perspectives.

Table 3. **Quantitative ablation study on the effectiveness of matching attention.** Temporal Attn denotes vanilla 1D temporal attention in SVD, while cross-frame Attn is inflated from spatial attention along temporal axis.

Attention Type	CLIP-FID ($\downarrow$)	LPIPS($\downarrow$)	SSIM ($\uparrow$)
Temporal Attn	8.398	0.165	0.807
Cross-Frame Attn	8.099	0.156	0.826
Matching Attn (Ours)	**7.783**	**0.140**	**0.859**

Table 4. **Quantitative ablation study on the effectiveness of source image reconstruction.** w/ Reconstruction means reconstructing both source and target images in diffusion loss.

Name	CLIP-FID ($\downarrow$)	LPIPS($\downarrow$)	SSIM ($\uparrow$)
w/o Reconstruction	8.201	0.154	0.834
w/ Reconstruction (Ours)	**7.783**	**0.140**	**0.859**

4.3. Emerging Capabilities of FramePainter

Despite being trained on only thousands of samples from real-world videos, FramePainter has showcased exceptional generalization across a wide range of scenarios, particularly in out-of-domain contexts. We provide several representative examples in Fig. 5. Firstly, FramePainter supports highly simplified editing instructions and allows users to conveniently modify images, *e.g.*, two sketches are enough to move the antennae or close the eyes in row 1. Secondly, it offers intuitive and precise control over complex editing signals, *e.g.*, adjust the shape of the flames and horse tail in row 2. Surprisingly, it is also capable of effectively handling

Figure 6. **Qualitative ablation study on the effectiveness of matching attention.** Matching attention obtains plausible edited results with fine-grained visual consistency. In contrast, temporal attention fails to handle editing signals involving large edited areas, while cross-frame attention struggles to precisely capture appearance.
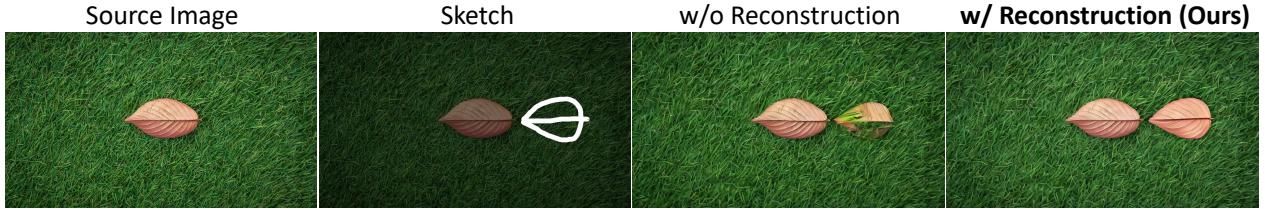

Figure 7. **Qualitative ablation study on source image reconstruction.** Compared to w/o reconstruction, reconstruction source image in diffusion loss can better preserve its color and texture and produce more visually consistent edited image.

Figure 8. **Visualization of attention weights and dense correspondence.** The attention map is computed between the selected target image token (*i.e.*, red query point) and all source image tokens. Among all source image tokens, the token with the highest similarity is marked as the matching point. We only visualize the tokens of foreground objects for simplicity.

out-of-domain scenarios (*i.e.*, cases not found in real-world videos). For example, in row 3, users can draw sketches to enlarge the beak of bird or transform the latte art into heart shape.

4.4. Ablation Study

Effectiveness of Matching Attention. Fig. 6 and Table 3 compares the performance of FramePainter using different attention mechanisms. Due to the limitation of receptive field, temporal attention fails to handle editing instructions involving large edited areas (*e.g.*, duplicate mushrooms). Albeit cross-frame attention can understand editing instructions involving significant changes, it struggles to make each target image token accurately query source image tokens (in Fig. 8), leading to the inconsistency in mushroom color. In contrast, matching attention improves the accuracy of queries and achieves promising dense correspondence between target and source images (*i.e.*, 3rd row in

Fig.. 8). Therefore, matching attention visibly exceeds both temporal and cross-frame attention in terms of quantitative metrics and visual consistency.

Effectiveness of Source Image Reconstruction. Existing image editing methods mainly focus on producing target images with diffusion loss, but ignore source image reconstruction. Our formulation allows the model to naturally reconstruct both target and source images. We investigate the impact of source image reconstruction in Fig. 7 and Table 4. From Fig. 7, performing reconstruction improves the visual coherence and realism of edited image (*e.g.*, the color and texture of leaf). Table 4 also highlights the effectiveness of source image reconstruction, *i.e.*, achieves lower CLIP-FID and LPIPS scores.

5. Conclusion

We reframe interactive image editing as an image-to-video generation task, and introduce FramePainter to leverage powerful priors of video diffusion models. Built upon Stable Video Diffusion, FramePainter greatly reduces training costs while ensuring seamless and coherent image edits. Considering the limitations of temporal attention, we propose matching attention to further improve visual consistency by ensuring dense correspondences between the edited and source images. The experiments highlights the effectiveness and efficiency of FramePainter, which achieves superior performance than previous state-of-the-art methods with far less training costs. Additionally, it demonstrates strong generalization to scenarios not present in real-world videos, such as transforming the clownfish into shark- like shape. We hope our work will inspire other image generative tasks that involve priors from videos.

References

[1] Hadi Alzayer, Zhihao Xia, Xuaner Zhang, Eli Shechtman, Jia-Bin Huang, and Michael Gharbi. Magic fixup: Streamlining photo editing by watching dynamic videos. *ArXiv*, abs/2403.13044, 2024. 2, 3, 4, 5

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A spacetime diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 2, 3

[4] 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. 2, 3, 4

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2, 3

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2

[7] 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. In *ICCV*, 2023. 2, 3

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2, 3

[9] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6593–6602, 2023. 3

[10] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023. 4

[11] Yutao Cui, Xiaotong Zhao, Guozhen Zhang, Shengming Cao, Kai Ma, and Limin Wang. Stabledrag: Stable dragging for point-based image editing. *ArXiv*, abs/2403.04437, 2024. 2, 3

[12] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *ICCV*, 2023. 2, 3

[13] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 4

[14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2025. 3

[15] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3

[16] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2, 3

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. 2, 3

[18] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 2

[19] Tianyu Huang, Yihan Zeng, Zhilu Zhang, Wan Xu, Hang Xu, Songcen Xu, Rynson WH Lau, and Wangmeng Zuo. Dreamcontrol: Control-based text-to-3d generation with 3d self-prior. *arXiv preprint arXiv:2312.06439*, 2023. 2

[20] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024. 3

[21] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 2, 5

[22] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023. 2, 3

[23] 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. 2, 4

[24] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022. 5

[25] Haofeng Liu, Chenshu Xu, Yifei Yang, Lihua Zeng, and Shengfeng He. Drag your noise: Interactive point-based editing via diffusion semantic propagation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6743–6752, 2024. 2, 3

[26] Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. In *European Conference on Computer Vision*, pages 1–17. Springer, 2025. 2

[27] Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8217–8227, 2023. 3

[28] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023. 2

[29] Zhengyao Lv, Yuxiang Wei, Wangmeng Zuo, and Kwan-Yee K. Wong. Place: Adaptive layout-semantic fusion for semantic image synthesis. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2

[30] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 2

[31] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Chenyang Qi, Chengfei Cai, Xiu Li, Zhifeng Li, Heung-Yeung Shum, Wei Liu, et al. Follow-your-click: Open-domain regional image animation via short prompts. *arXiv preprint arXiv:2403.08268*, 2024. 3

[32] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2, 3

[33] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 2

[34] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 2, 3

[35] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 2, 3

[36] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. *arXiv preprint arXiv:2402.02583*, 2023. 12

[37] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 2, 3

[38] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 2

[39] 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*, 2022. 2

[40] Shen Nie, Hanzhong Allan Guo, Cheng Lu, Yuhao Zhou, Chenyu Zheng, and Chongxuan Li. The blessing of randomness: Sde beats ode in general diffusion-based image editing. *arXiv preprint arXiv:2311.01410*, 2023. 3, 5, 12

[41] Xingang Pan, Ayush Kumar Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 2, 3

[42] 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. 3

[43] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 2, 3

[44] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 4

[45] 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. 1, 2

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[49] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 1, 2

[50] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8839–8849, 2023. 2, 3, 12

[51] Yujun Shi, Jun Hao Liew, Hanshu Yan, Vincent Y. F. Tan, and Jiashi Feng. Lightningdrag: Lightning fast and accurate drag-based image editing emerging from videos. In *arXiv preprint arXiv:2405.13722*, 2024. 2, 3, 4, 5, 12

[52] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 4

[53] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2024. 3

[54] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 2, 3

[55] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2, 3

[56] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pages 36–54. Springer, 2025. 5

[57] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2

[58] Yuxiang Wei, Zhilong Ji, Jinfeng Bai, Hongzhi Zhang, Lei Zhang, and Wangmeng Zuo. Masterweaver: Taming editability and face identity for personalized text-to-image generation. In *European Conference on Computer Vision*, 2024. 2

[59] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024. 3

[60] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2

[61] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 2, 3

[62] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2

[63] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. *arXiv preprint arXiv:2311.10982*, 2023. 2, 3

[64] Xiaohang Zhan, Xingang Pan, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised learning via conditional motion propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1881–1889, 2019. 5

[65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[66] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[67] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2, 3

[68] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 2, 4

[69] Yabo Zhang, Yuxiang Wei, Xianhui Lin, Zheng Hui, Peiran Ren, Xuansong Xie, Xiangyang Ji, and Wangmeng Zuo. Videoelevator: Elevating video generation quality with versatile text-to-image diffusion models. *arXiv preprint arXiv:2403.05438*, 2024. 2

[70] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2024. 2

[71] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2, 3

A. Implementation Details of Different Visual Editing Instructions.

By default, the visual editing instructions (*e.g.*, sketch images and coarsely edited images) are directly encoded using sparse control encoder and injected into the denoising U-Net. However, it is challenging to encode images that only contain source and target points, which cannot accurately represent the correspondence between each pair of points. Since this paper aims to explore a general paradigm for interactive learning, rather than focusing on the specific editing method of dragging points, we adopt a simple and intuitive way to encode dragging points. Specifically, at the output of each attention block, we directly copy the source image tokens corresponding to the positions of source points, and add them to the edited image tokens at the positions of target points. As a result, this simple approach allows for an accurate understanding of dragging points and enables plausible editing of input images, *e.g.*, in Fig. 4 and Fig. 12.

B. More Visualizations and Comparisons.

Fig. 9 show more visualizations on sketch images. Fig. 10 and Fig. 11 provide more comparisons with alternative approaches on sketch images and coarsely edited images, respectively. Fig. 12 compares a wide range of drag-based methods, including encoder-based (*i.e.*, LightningDrag [51]) and optimization-based (*i.e.*, DragDiffusion [50], SDE-Drag [40], and DiffEditor [36]. Compared to the baselines, FramePainter presents superior performance in understanding the dragging points and maintaining the structural integrity of objects. In contrast, due to the absence of real-world dynamic priors, optimization-based methods struggle with moving object parts, *e.g.*, duplicated tail in row 2 of Fig. 12 (top) and duplicate hair in row 2 of Fig. 12 (bottom). Encoder-based method cannot preserve the overall structure of objects, *e.g.*, separated mushrooms in row 1 of Fig. 12 (top).
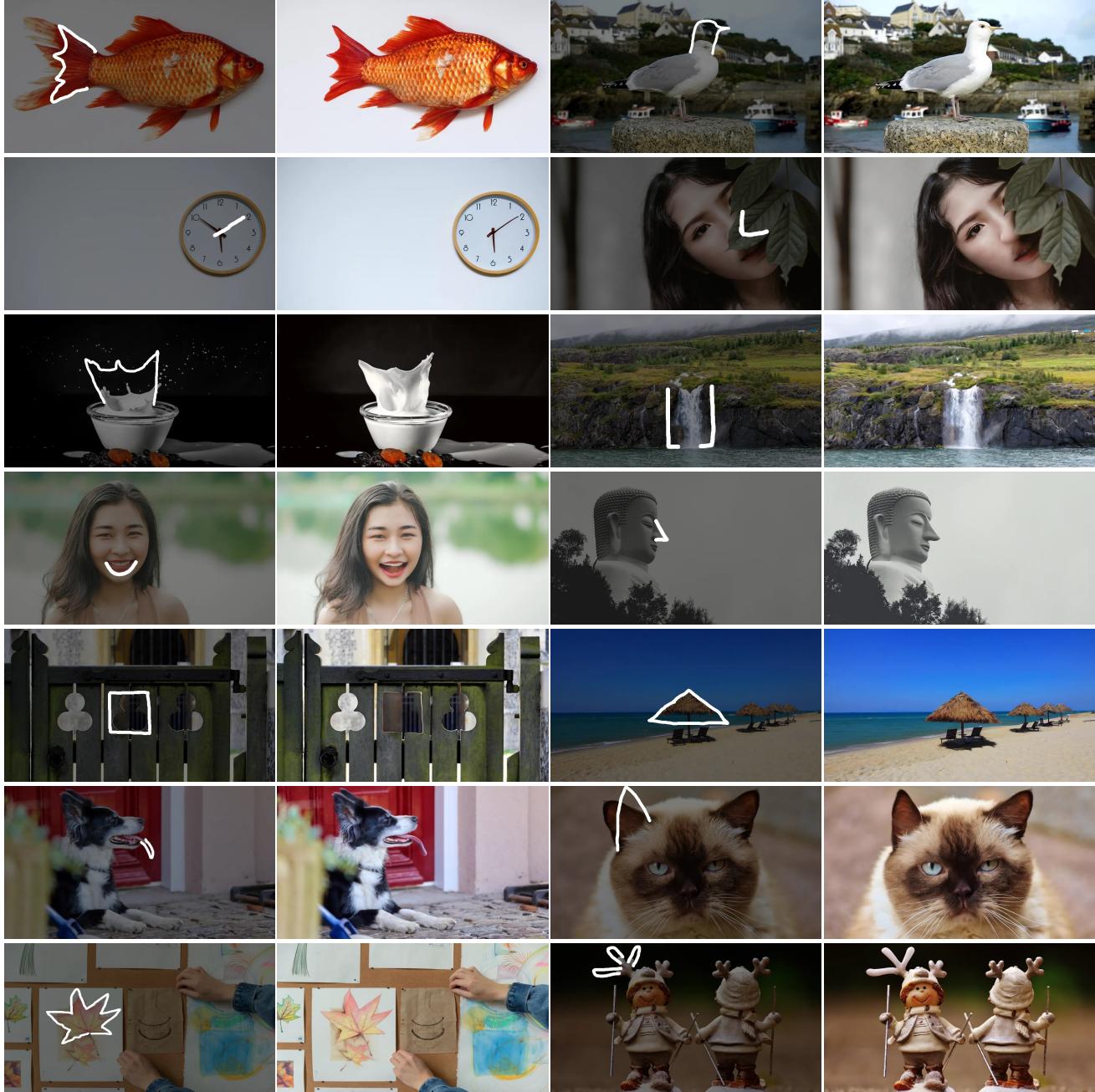
Figure 9. **More visualization examples of FramePainter.** This figure presents both a wide range of scenarios, including in-domain (*e.g.*, change the position of cat ear) and out-of-domain cases (*e.g.*, enlarge the dear horn in hat).
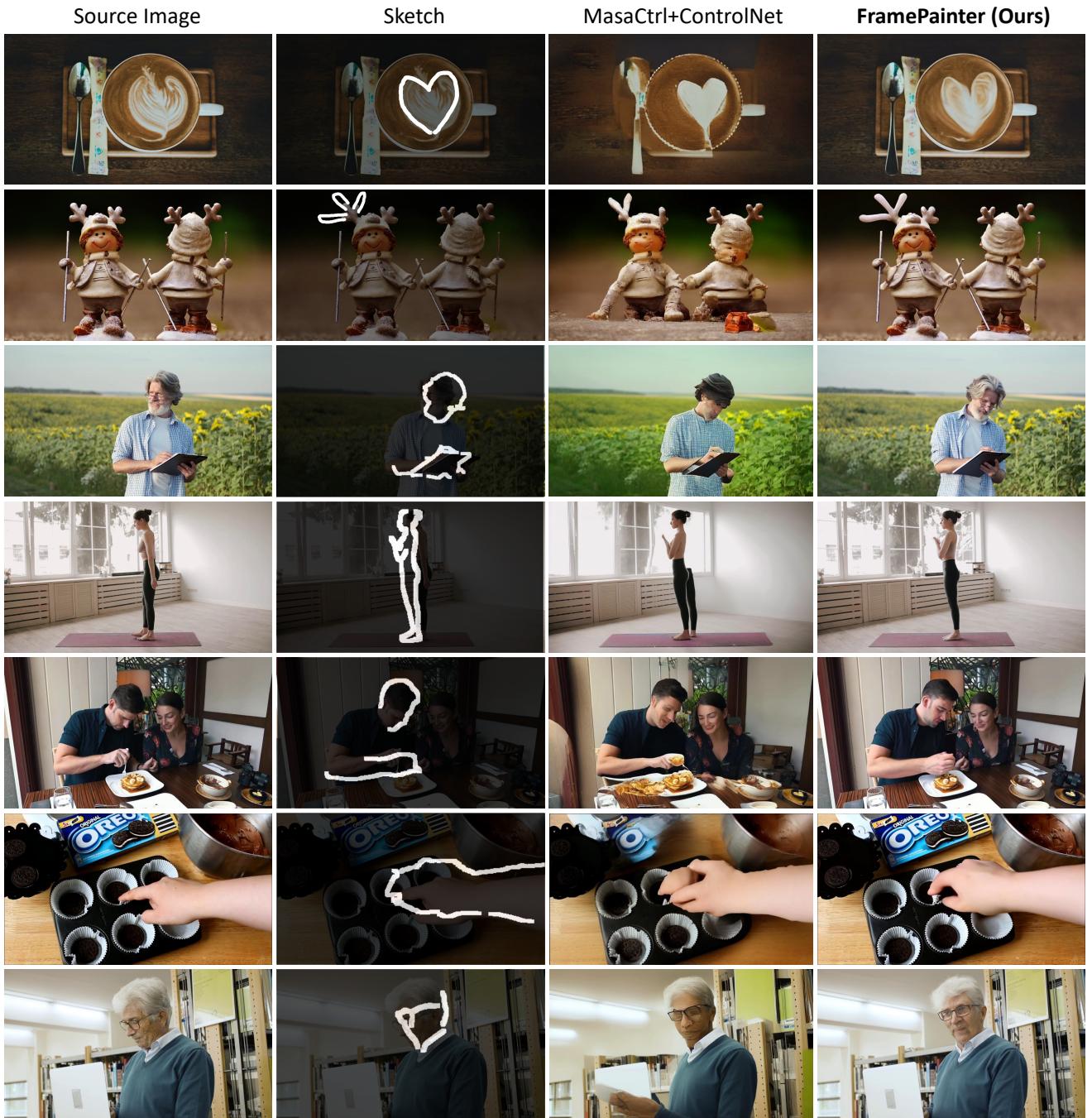
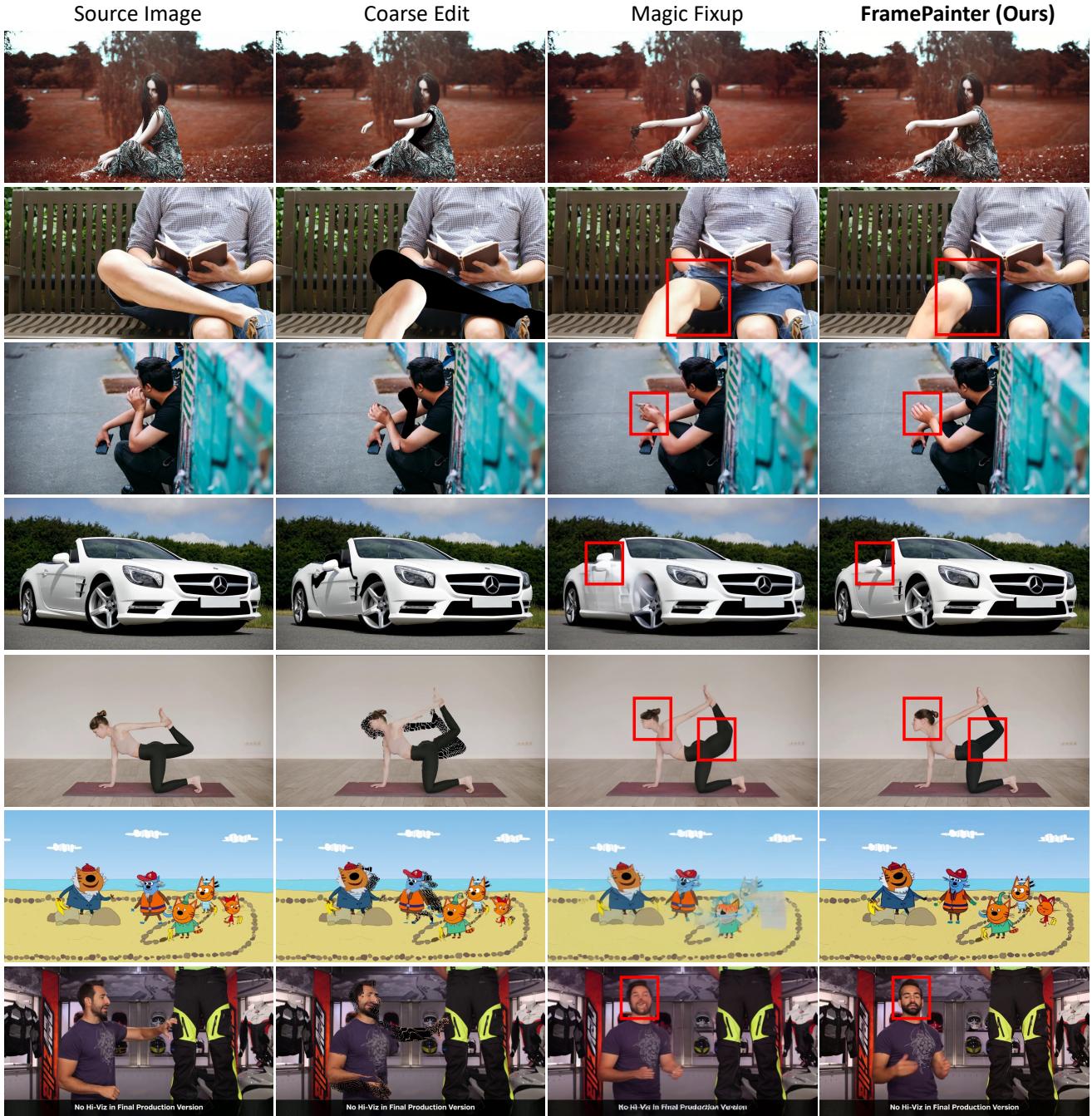Figure 10. **More qualitative comparisons in sketch images.**

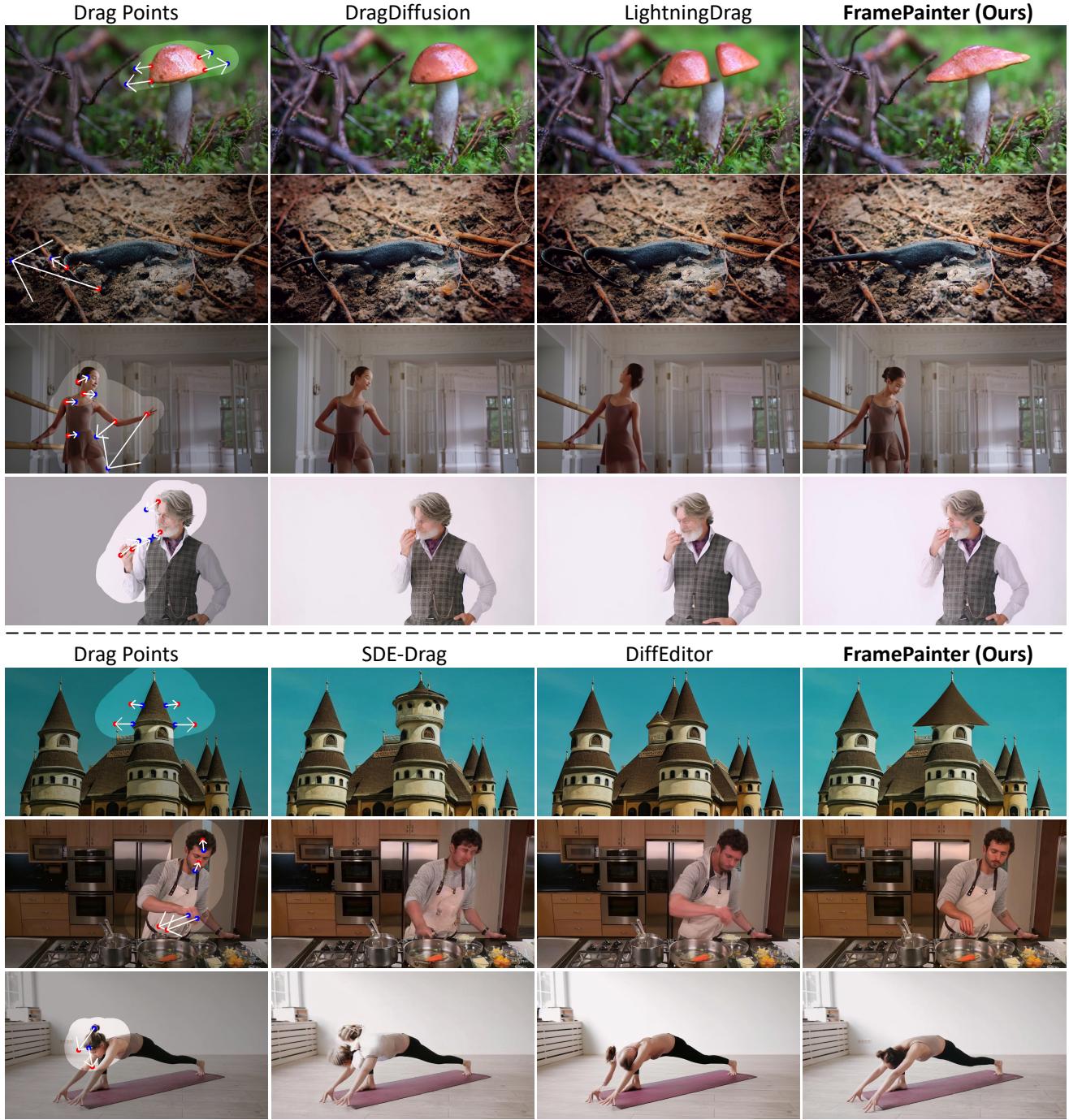Figure 11. **More qualitative comparisons in coarsely edited images.**

Figure 12. **More qualitative comparisons in dragging points.** We compare FramePainter with both encoder-based (*i.e.*, LightningDrag) and optimization-based methods (*i.e.*, DragDiffusion, SDE-Drag, and DiffEditor). During inference, DragDiffusion and SDE-Drag require to finetune additional LoRAs to preserve the visual appearance of source images.