VideoPainter: Any-length Video Inpainting and Editing with Plug-and-Play Context Control

YUXUAN BIAN, The Chinese University of Hong Kong, China
ZHAOYANG ZHANG*, Tencent ARC Lab, China
XUAN JU, The Chinese University of Hong Kong, China
MINGDENG CAO, The University of Tokyo, Japan
LIANGBIN XIE, University of Macau, China
YING SHAN, Tencent ARC Lab, China
QIANG XU†, The Chinese University of Hong Kong, China

Fig. 1. ***VideoPainter* enables plug-and-play text-guided video inpainting and editing for any video length and pre-trained Diffusion Transformer with masked video and video caption (user editing instruction).** The upper part demonstrates the effectiveness of *VideoPainter* in various video inpainting scenarios, including object, landscape, human, animal, multi-region (Multi), and random masks. The lower section demonstrates the performance of *VideoPainter* in video editing, including adding, removing, changing attributes, and swapping objects. In both video inpainting and editing, we demonstrate strong ID consistency in generating long videos (Any Len.). **Project page for this paper is at: https://yxbian23.github.io/project/video-painter**

Video inpainting, crucial for the media industry, aims to restore corrupted content. However, current methods relying on limited pixel propagation or single-branch image inpainting architectures face challenges with generating fully masked objects, balancing background preservation with foreground generation, and maintaining ID consistency over long video. To address these issues, we propose *VideoPainter*, an efficient dual-branch framework featuring a lightweight context encoder. This plug-and-play encoder processes masked videos and injects background guidance into any pre-trained video diffusion transformer, generalizing across arbitrary mask types, enhancing background integration and foreground generation, and enabling

*Project Lead.
†Corresponding Author.

Authors' addresses: Yuxuan Bian, The Chinese University of Hong Kong, Hong Kong, China, yuxuanbian23@gmail.com; Zhaoyang Zhang, Tencent ARC Lab, Shenzhen, China, zhaoyangzhang@link.cuhk.edu.hk; Xuan Ju, The Chinese University of Hong Kong, Hong Kong, China, juxuan.27@gmail.com; Mingdeng Cao, The University of Tokyo, Tokyo, Japan, mingdengcao@gmail.com; Liangbin Xie, University of Macau, Macau, China, lb.xie@siat.ac.cn; Ying Shan, Tencent ARC Lab, Shenzhen, China, yingsshan@tencent.com; Qiang Xu, The Chinese University of Hong Kong, Hong Kong, China, qxu@cse.cuhk.edu.hk.

user-customized control. We further introduce a strategy to resample inpainting regions for maintaining ID consistency in any-length video inpainting. Additionally, we develop a scalable dataset pipeline using advanced vision models and construct *VPData* and *VPBench*—the largest video inpainting dataset with segmentation masks and dense caption (>390K clips) —to support large-scale training and evaluation. We also show *VideoPainter*'s promising potential in downstream applications such as video editing. Extensive experiments demonstrate *VideoPainter*'s state-of-the-art performance in any-length video inpainting and editing across 8 key metrics, including video quality, mask region preservation, and textual coherence.

CCS Concepts: • **Computing methodologies → Computer vision**.

Additional Key Words and Phrases: Artificial Intelligence Generative Content, Computer Vision, Video Inpainting, Video Editing

ACM Reference Format:
Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. 2025. VideoPainter: Any-length Video Inpainting and Editing with Plug-and-Play Context Control. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25), August 10–14, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3721238.3730673

1 INTRODUCTION

Video inpainting [Quan et al. 2024], which aims to restore the corrupted video while maintaining coherence, facilitates numerous applications, including try-on [Fang et al. 2024], film production [Polyak et al. 2024], and video editing [Sun et al. 2024]. Recently, Diffusion Transformers (DiT) [OpenAI 2024; Peebles and Xie 2023] have shown promise in video generation, leading to the exploration of generative video inpainting [Zhang et al. 2024b; Zi et al. 2024].

Existing approaches, as illustrated in Fig. 2, can be broadly categorized into two types: (1) Non-Generative methods [Lee et al. 2019; Li et al. 2022; Zhou et al. 2023] depend on limited pixel feature propagation (physical constraints or model architectural priors), which only take masked videos as inputs and cannot generate fully segmentation-masked objects. (2) Generative methods [Wang et al. 2024; Zhang et al. 2024b; Zi et al. 2024] extend single-branch image inpainting architectures [Rombach and Esser 2022] to video by incorporating temporal attention, which struggles to balance background preservation and foreground generation in one model and obtain inferior temporal coherence compared to native video DiTs. Moreover, both paradigms neglect long video inpainting and struggle to maintain consistent object identity with long videos.

This motivates us to decompose video inpainting into background preservation and foreground generation and adopt a dual-branch architecture in DiTs, where we can incorporate a dedicated context encoder for masked video feature extraction while utilizing the pre-trained DiT's capabilities to generate semantic coherent video content conditioned on both the preserved background and text prompts. Similar observations have been made in image inpainting research, notably in BrushNet [Ju et al. 2024] and ControlNet [Zhang et al. 2023]. However, directly applying their architecture to video DiTs presents several challenges: (1) Given Video DiT's robust generative foundation and heavy model size, replicating the full/half-giant Video DiT backbone as the context encoder would be unnecessary and computationally prohibitive. (2) Unlike BrushNet's pure convolutional control branch, DiT's tokens in masked regions inherently contain background information due to global attention, complicating the distinction between masked and unmasked regions in DiT backbones. (3) ControlNet lacks feature injection across all layers, hindering dense background control for inpainting tasks.

To address these challenges, we introduce *VideoPainter*, which enhances pre-trained DiT with a lightweight context encoder comprising only 6% of the backbone parameters, to form the first efficient dual-branch video inpainting architecture. *VideoPainter* features three main components: (1) A streamlined context encoder with just two layers, which integrates context features into the pre-trained DiT in a group-wise manner, ensuring efficient and dense background guidance. (2) Mask-selective feature integration to clearly distinguish the tokens of the masked and unmasked region. (3) A novel inpainting region ID resampling technique to efficiently process videos of any length while maintaining ID coherence. By freezing the pre-trained context encoder and DiT backbone, and adding an ID-Adapter, we enhance the backbone's attention sampling by concatenating the original key-value vectors with the inpainting region tokens. During inference, inpainting region tokens from previous clips are appended to the current key-value vectors, ensuring the long-term preservation of target IDs. Notably, our *VideoPainter* supports plug-and-play and user-customized control.

For large-scale training, we develop a scalable dataset pipeline using advanced vision models [OpenAI 2024; Ravi et al. 2024; Zhang et al. 2024a], constructing the largest video inpainting dataset, *VPData*, and benchmark, *VPBench*, with over 390K clips featuring precise segmentation masks and dense text captions. We further demonstrate *VideoPainter*'s potential by establishing an inpainting-based video editing pipeline that delivers promising results.

To validate our approach, we compare *VideoPainter* against previous state-of-the-art (SOTA) baselines and a single-branch finetuning setup that combines noisy latent, masked video latent, and mask at the input channel. *VideoPainter* demonstrates superior performance in both training efficiency and final results.

In summary, our contributions are as follows:

- We propose **VideoPainter**, the first dual-branch video inpainting framework that supports plug-and-play background controls.
- We design a lightweight context encoder for efficient and dense background control, and inpainting region ID resampling for ID consistency in any-length video inpainting and editing.
- We introduce **VPData**, the largest video inpainting datasets comprising over 390K clips (> 866.7 hours), and **VPBench**, both featuring precise masks and detailed video captions.
- Experiments show *VideoPainter* achieves state-of-the-art performance across 8 metrics including video quality, masked region preservation, and text alignment in video inpainting and editing.

2 RELATED WORK

2.1 Video Inpainting

Video inpainting approaches can be broadly classified into two categories based on whether they possess generative capabilities:

Non-generative methods. These methods [Hu et al. 2020; Li et al. 2022; Zhang et al. 2022a,b; Zhou et al. 2023] leverage architecture priors to facilitate pixel propagation. This includes utilizing local

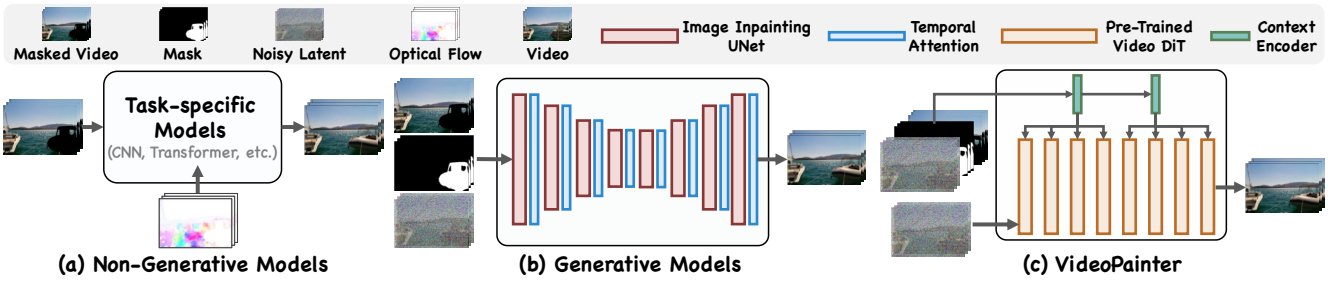

Fig. 2. Framework Comparison. Non-generative approaches, limited to pixel propagation from backgrounds, fail to inpaint fully segmentation-masked objects. Generative methods adapt single-branch image inpainting models to video by adding temporal attention, struggling to maintain background fidelity and generate foreground contents in one model. In contrast, *VideoPainter* implements a dual-branch architecture that leverages an efficient context encoder with any pre-trained DiT, decoupling video inpainting to background preservation and foreground generation, and enabling plug-and-play video inpainting control.

perception of 3D CNNs [Chang et al. 2019a,b; Hu et al. 2020; Wang et al. 2019], and exploiting the global perception of attention to retrieve and aggregate tokens with similar texture for filling masked video [Lee et al. 2019; Liu et al. 2021; Zeng et al. 2020; Zhang et al. 2022a]. They also introduce various physical quantities, especially optical flow, as auxiliary conditions as it simplifies RGB pixel inpainting by completing less complex flow fields [Gao et al. 2020; Kim et al. 2019; Li et al. 2020; Xu et al. 2019; Zhang et al. 2022b,c; Zou et al. 2021]. However, they are only effective for partial object occlusions with random masks but face significant limitations when inpaint fully masked regions due to insufficient contexts.

Generative methods. Recent advances in generative foundation models [Guo et al. 2023; Rombach et al. 2022] have sparked numerous approaches that leverage additional modules or training strategies to extend backbones' capabilities for video inpainting [Wang et al. 2024; Zhang et al. 2024b; Zi et al. 2024]. AVID [Zhang et al. 2024b] and COCOCO [Zi et al. 2024] represent the most related recent works. Both adopt a similar implementation by augmenting Stable Diffusion Inpainting [Rombach and Esser 2022] with trainable temporal attention layers. This architecture includes per-frame region filling based on the image inpainting backbone and temporal smoothing with temporal attention. Despite showing promising results for both random and segmentation masks due to their generative abilities, they struggle to balance background preservation and foreground generation with text caption [Ju et al. 2024; Li et al. 2024] within the single backbone. AVID also explores any-length video inpainting by smoothing latent at segment boundaries and using the middle frame as the ID reference. In contrast, *VideoPainter* is a dual-branch framework by decoupling video inpainting into foreground generation and background-guided preservation. It employs an efficient context encoder to guide any pre-trained DiT, facilitating plug-and-play control. Furthermore, *VideoPainter* also introduces a novel inpainting region ID resampling technique that enables ID consistency in any-length video inpainting.

2.2 Video Inpainting Datasets

Recent advances in segmentation [Ravi et al. 2024] have created many video segmentation datasets [Darkhalil et al. 2022; Ding et al. 2023; Hong et al. 2023; Perazzi et al. 2016; Tokmakov et al. 2023; Xu et al. 2018]. Among these, DAVIS [Perazzi et al. 2016] and YouTube-VOS [Xu et al. 2018] have become prominent benchmarks for video inpainting due to their high-quality masks and diverse object categories. However, the existing datasets face two primary limitations: (1) insufficient scale to meet the data requirements of generative models, and (2) the absence of crucial control conditions necessary for generating masked objects such as video captions. In contrast, as shown in Tab. 1, we developed a scalable dataset pipeline based on state-of-the-art vision understanding models [OpenAI 2024; Ravi et al. 2024; Zhang et al. 2024a], and constructed the largest video inpainting dataset to date with over 390K clips, each annotated with segmentation masks and dense video captions.

Table 1. Comparison of video inpainting datasets. Our *VPData* is the largest video inpainting dataset to date, comprising over 390K high-quality clips with segmentation masks, video captions, and masked region descriptions.

Dataset	#Clips	Duration	Video Caption	Masked Region Desc.
DAVIS [Perazzi et al. 2016]	0.4K	0.1h	✗	✗
YouTube-VOS [Xu et al. 2018]	4.5K	5.6h	✗	✗
VOST [Tokmakov et al. 2023]	1.5K	4.2h	✗	✗
MOSE [Ding et al. 2023]	5.2K	7.4h	✗	✗
LVOS [Hong et al. 2023]	1.0K	18.9h	✗	✗
SA-V [Ravi et al. 2024]	642.6K	196.0h	✗	✗
Ours	390.3K	866.7h	✓	✓

3 METHOD

Sec. 3.1 and Fig. 3 illustrate our pipeline for building *VPData* and *VPBench*. Sec. 3.2 and Fig. 4 show our dual-branch *VideoPainter*. Sec. 3.3 and Sec. 3.4 introduce our inpainting region ID resampling approach for any-length video inpainting and plug-and-play control.

3.1 *VPData* and *VPBench* Construction Pipeline

To address the challenges of limited size and lack of text annotations, we present a scalable dataset pipeline leveraging advanced vision models [OpenAI 2024; Ravi et al. 2024; Zhang et al. 2024a]. This leads to *VPData* and *VPBench*, the largest video inpainting dataset and benchmark with precise masks and video/masked region captions. As shown in Fig. 3, the pipeline involves 5 steps: collection, annotation, splitting, selection, and captioning.

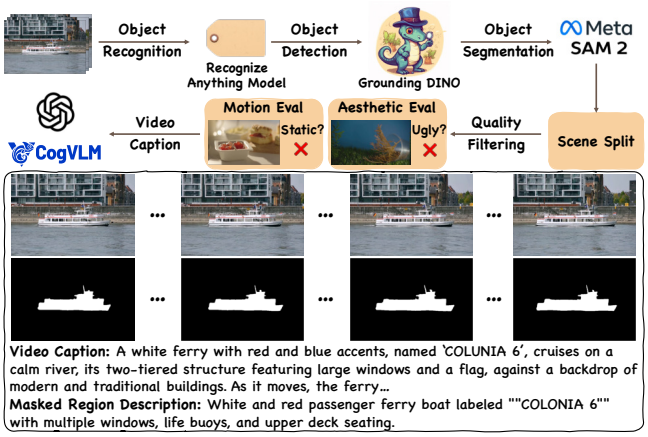

Fig. 3. Dataset Construction Pipeline. It consists of five pre-processing steps: collection, annotation, splitting, selection, and captioning.

Collection. We chose Videvo and Pexels [1] as our data sources. We finally obtained around 450K videos from these sources.

Annotation. For each collected video, we implement a cascaded workflow for automated annotation:

- We employ the Recognize Anything Model [Zhang et al. 2024a] for open-set video tagging to identify primary objects.
- Based on the detected object tags, we utilize Grounding DINO [Liu et al. 2023] to detect bounding boxes for objects at fixed intervals.
- These bounding boxes serve as prompts for SAM2 [Ravi et al. 2024], which generates high-quality mask segmentations.
- Then we employ rigorous filtering criteria: inter-frame mask area variation $\Delta < 20\%$ and frame coverage maintained between 30% – 70% to ensure reliable segmentation masks quality.

Splitting. Scene transitions may occur while tracking the same object from different angles, causing disruptive view changes. We utilize PySceneDetect [Castellano 2024] to identify scene transitions and subsequently partition the masks. Then we segmented the sequences into 10-second intervals and discarded short clips (< 6s).

Selection. We employ 3 key criteria: (1) Aesthetic Quality, evaluated using the Laion-Aesthetic Score Predictor [Schuhmann et al. 2022]; (2) Motion Strength, predicted by optical flow measurements using the RAFT[Teed and Deng 2020]; and (3) Content Safety, assessed via the Stable Diffusion Safety Checker [Rombach et al. 2022].

Captioning. As Tab. 1 shows, existing video segmentation datasets lack textual annotations, primary conditions in generation [Betker et al. 2023; Chen et al. 2023], creating a data bottleneck for applying generative models to video inpainting. Therefore, we leverage SOTA VLMs, specifically CogVLM2 [Wang et al. 2023] and GPT-4o [OpenAI 2024], to uniformly sample keyframes and generate dense video captions and detailed descriptions of the masked objects.

3.2 Dual-branch Inpainting Control

We incorporate masked video features into the pre-trained diffusion transformer (DiT) via an efficient context encoder, to decouple the background context extraction and foreground generation. This encoder processes a concatenated input of noisy latent, masked video latent, and downsampled masks. Specifically, the noisy latent provides information about the current generation. The masked video latent, extracted via VAE, aligns with the pre-trained DiT's latent distribution. We apply cubic interpolation to downsample masks, ensuring dimensional compatibility between masks and latents.

Based on DiT's inherent generative abilities [OpenAI 2024], the control branch only needs to extract contextual cues to guide the backbone in preserving background and generating foreground. Therefore, instead of previous heavy approaches that duplicate half or all of the backbone [Ju et al. 2024; Zhang et al. 2023], *VideoPainter* employs a lightweight design by cloning only the first two layers of pre-trained DiT, accounting for merely 6% of the backbone parameters. The pre-trained DiT weights provide a robust prior for extracting masked video features. The context encoder features are integrated into the frozen DiT in a group-wise, token-selective manner. The group-wise feature integration is formulated as follows: the first layer's features are added back to the initial half of the backbone, while the second layer's features are integrated into the latter half, achieving lightweight and efficient context control. The token-selective mechanism is a pre-filtering process, where only tokens representing pure background are added back, while others are excluded from integration, as shown in the upper right of Fig. 4. This ensures that only the background context is fused into the backbone, preventing potential ambiguity during backbone generation.

The feature integration is shown in Eq. 1. $\epsilon_\theta\left(z_t, t, C\right)_i$ indicates the feature of the i-th layer in DiT ϵ_θ with $i \sim [1, n]$, where n is the number of layers. The same notation applies to $\epsilon_\theta^{VideoPainter}$, which takes the concatenated noisy latent z_t, masked video latent z_0^{masked}, and downsampled mask $m^{resized}$ as input. The concatenation operation is denoted as $[\cdot]$. $\mathcal{Z}$ is the zero linear operation.

$$\epsilon_\theta\left(z_t, t, C\right)_i = \epsilon_\theta\left(z_t, t, C\right)_i + \mathcal{Z}\left(\epsilon_\theta^{VideoPainter}\left(\left[z_t, z_0^{masked}, m^{resized}\right], t\right)_{i//\frac{n}{2}}\right) \quad (1)$$

3.3 Target Region ID Resampling

While current DiTs show promise in handling temporal dynamics [Bian et al. 2024; Kuaishou 2024], they struggle to maintain smooth transitions and long-term identity consistency.

Smooth Transition. Following AVID [Zhang et al. 2024b], we employ overlapping generation and weighted average to maintain consistent transitions. Additionally, we utilize the last frame of the previous clip (before overlap) as the first frame of the current clip's overlapping region to ensure visual appearance continuity.

Identity Consistency. To maintain identity consistency in the long video, we introduce an inpainting region ID resampling method, as shown in lower Fig. 4. During training, we freeze both the DiT and the context encoder. Then we add trainable ID-Resample Adapters into the frozen DiT (LoRA), enabling ID resampling functionality. Specifically, tokens from the current masked region, which contain the desired ID, are concatenated with the KV vectors, thereby enhancing ID preservation in the inpainting region through additional KV resampling. During inference, we prioritize maintaining ID consistency with the inpainting region tokens from the previous clip, as it represents the most temporally proximate generated result. Specifically, given current Q_i^v, K_i^v, and V_i^v, we concatenate tokens

[1] Videvo: https://www.videvo.net/, Pexels: https://www.pexels.com/

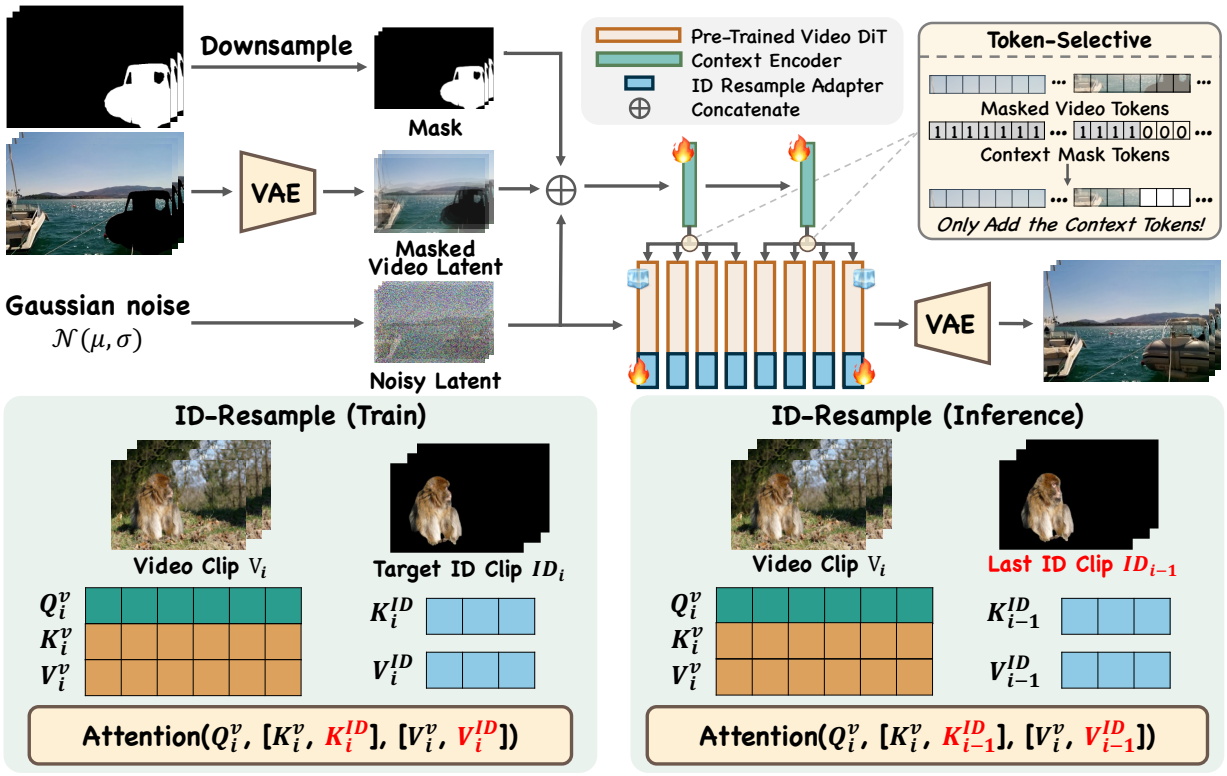

Fig. 4. **Model overview. The upper figure** shows the architecture of *VideoPainter*. The context encoder performs video inpainting based on concatenation of the noisy latent, downsampled masks, and masked video latent via VAE. Features extracted by the context encoder are integrated into the pre-trained DiT in a group-wise and token-selective manner, where two encoder layers modulate the first and second halves of the DiT, respectively, and only the background tokens will be integrated into the backbone to prevent information ambiguity. **The lower figure** illustrates the inpainting ID region resampling with the ID Resample Adapter. During training, tokens of the current masked region are concatenated to the KV vectors, enhancing ID preservation of the inpainting region. During inference, the ID tokens of the last clip are concatenated to the current KV vectors, maintaining ID consistency with the last clip by resampling.

containing ID information (K_i^{id} and V_i^{id}) to the current KV pairs (during training, these are tokens from the current inpainting region; during inference, from the previous clip's inpainting region). This forms new KV-vectors $[K_i^v, K_i^{id}]$ and $[V_i^v, V_i^{id}]$ (where $[\cdot,\cdot]$ denotes concatenation), enabling the model to sample necessary ID information and better maintain ID consistency.

3.4 Plug-and-Play Control

Our plug-and-play framework demonstrates versatility across two aspects: it supports various stylization backbones or LoRAs and is compatible with both text-to-video (T2V) [NVIDIA 2025; Yang et al. 2024] and image-to-video (I2V) [Guo et al. 2024; Shi et al. 2024] DiT architectures. The I2V compatibility particularly enables seamless integration with existing image inpainting capabilities. When utilizing an I2V DiT backbone, *VideoPainter* requires only one additional step: generating the initial frame using any image inpainting model guided by the masked region's text caption. This inpainted frame then serves as both the image condition and the first masked video frame. These capabilities further demonstrate the exceptional transferability and versatility of *VideoPainter*.

4 EXPERIMENTS

4.1 Implementation details

VideoPainter is built upon a pre-trained Image-to-Video Diffusion Transformer CogVideo-5B-I2V [Yang et al. 2024] (by default) and its Text-to-Video version. In training, we use *VPData* at a 480×720 resolution, learning rate 1×10^{-5}, batch size 1 for both the context encoder (80,000 steps) and the ID Resample Adapter (2,000 steps) in two stages with AdamW. In training, we randomly sample dilation and erosion with kernel sizes $\in [8, 32]$ to enhance robustness to mask precision. This also enables our random-mask inpainting.

Benchmarks. In video inpainting, we employ Davis [Perazzi et al. 2016] as the benchmark for random masks and *VPBench* for segmentation-based masks. *VPBench* consists of 100 6-second videos for standard video inpainting, and 16 videos with an average duration of more than 30 seconds for long video inpainting. The *VPBench* includes diverse content including objects, humans, animals, landscapes, and multi-range masks. For video editing evaluation, we also utilize *VPBench*, which includes four fundamental editing operations (add, remove, swap, and change) and comprises 45 6-second videos and 9 videos with an average duration of 30 seconds.

Metrics. We consider 8 metrics from three aspects: masked region preservation, text alignment, and video generation quality.

- *Masked Region Preservation.* We follow previous works using standard PSNR [Wikipedia contributors 2024c], LPIPS [Zhang et al. 2018], SSIM [Wang et al. 2004], MSE [Wikipedia contributors 2024b] and MAE [Wikipedia contributors 2024a] in the unmasked region among the generated video and the original video.
- *Text Alignment.* We employ CLIP Similarity (CLIP Sim) [Wu et al. 2021] to assess the semantic consistency between the generated video and its corresponding text caption. We also measure CLIP Similarity within the masked regions (CLIP Sim (M)).
- *Video Generation Quality.* Following previous methods, we use FVID [Wang et al. 2018] to measure the generated video quality.

4.2 Video Inpainting

Quantitative comparisons. Tab. 2 shows the quantitative comparison on *VPBench* and Davis [Perazzi et al. 2016]. We compare the inpainting results of non-generative ProPainter [Zhou et al. 2023], generative COCOCO [Zi et al. 2024], and Cog-Inp [Yang et al. 2024], a strong baseline proposed by us, which inpaint first frame using image inpainting models and use the I2V backbone to propagate results with the latent blending operation [Avrahami et al. 2023]. In the segmentation-based *VPBench*, ProPainter, and COCOCO exhibit the worst performance across most metrics, primarily due to the inability to inpaint fully masked objects and the single-backbone architecture's difficulty in balancing the competing background preservation and foreground generation, respectively. In the random mask benchmark Davis, ProPainter shows improvement by leveraging partial background information. However, *VideoPainter* achieves optimal performance across segmentation (standard and long length) and random masks through its dual-branch architecture that effectively decouples background preservation and foreground generation.

Qualitative comparisons. The qualitative comparison with previous video inpainting methods is shown in Fig. 5. *VideoPainter* consistently shows exceptional results in video coherence, quality, and alignment with text caption. Notably, ProPainter fails to generate fully masked objects because it only depends on background pixel propagation instead of generating. While COCOCO demonstrates basic functionality, it fails to maintain consistent ID in inpainted regions (inconsistent vessel appearances and abrupt terrain changes) due to its single-backbone architecture attempting to balance background preservation and foreground generation. Cog-Inp achieves basic inpainting results; however, its blending operation's inability to detect mask boundaries leads to significant artifacts. Moreover, *VideoPainter* can generate coherent videos exceeding one minute while maintaining ID consistency through our ID resampling.

4.3 Video Editing

VideoPainter can be used for video inpainting by employing Vison Language Models [OpenAI 2024; Team et al. 2024] to generate modified captions based on user editing instructions and source captions and apply *VideoPainter* to inpaint based on the modified captions. Tab. 3 shows the quantitative comparison on *VPBench*. We compare the editing results of inverse-based UniEdit [Bai et al. 2024], DiT-based DiTCtrl [Cai et al. 2024], and end-to-end ReVideo [Mou et al. 2024]. For both standard and long videos in *VPBench*, *VideoPainter* achieves superior performance, even surpassing the end-to-end ReVideo. This success can be attributed to its dual-branch architecture, which ensures excellent background preservation and foreground generation capabilities, maintaining high fidelity in non-edited regions while ensuring edited regions closely align with editing instructions, complemented by inpainting region ID resampling that maintains ID consistency in long video. The qualitative comparison with previous video inpainting methods is shown in Fig. 5. *VideoPainter* demonstrates superior performance in preserving visual fidelity and text-prompt consistency.

Table 2. **Quantitative comparisons among *VideoPainter* and other video inpainting models in *VPBench* for segmentation mask (Standard (S) and Long (L) Video) and Davis for random mask**: ProPainter [Zhou et al. 2023], COCOCO [Zi et al. 2024], and Cog-Inp [Yang et al. 2024]. Metrics include masked region preservation, text alignment, and video quality. **Red** stands for the best, **Blue** stands for the second best.

Metrics		Masked Region Preservation					Text Alignment		Video Quality
	Models	PSNR↑	SSIM↑	LPIPS$_{\times 10^2}$↓	MSE$_{\times 10^2}$↓	MAE$_{\times 10^2}$↓	CLIP Sim↑	CLIP Sim (M)↑	FVID↓
VPBench-S	ProPainter	20.97	**0.87**	9.89	1.24	**3.56**	7.31	17.18	0.44
	COCOCO	19.27	0.67	14.80	1.62	6.38	7.95	20.03	0.69
	Cog-Inp	**22.15**	0.82	**9.56**	**0.88**	3.92	**8.41**	**21.27**	**0.18**
	Ours	**23.32**	**0.89**	**6.85**	**0.82**	**2.62**	**8.66**	**21.49**	**0.15**
VPBench-L	ProPainter	**20.11**	**0.84**	**11.18**	**1.17**	**3.71**	9.44	17.68	0.48
	COCOCO	19.51	0.66	16.17	1.29	6.02	11.00	20.42	0.62
	Cog-Inp	19.78	0.73	12.53	1.33	5.13	**11.47**	**21.22**	**0.21**
	Ours	**22.19**	**0.85**	**9.14**	**0.71**	**2.92**	**11.52**	**21.54**	**0.17**
Davis	ProPainter	**23.99**	**0.92**	**5.86**	0.98	**2.48**	**7.54**	16.69	**0.12**
	COCOCO	21.34	0.66	10.51	0.92	4.99	6.73	17.50	0.33
	Cog-Inp	23.92	0.79	10.78	**0.47**	3.23	7.03	**17.53**	0.17
	Ours	**25.27**	**0.94**	**4.29**	**0.45**	**1.41**	**7.21**	**18.46**	**0.09**

Table 3. **Quantitative comparisons among *VideoPainter* and other video editing models in *VPBench* (Standard and Long Video)**: UniEdit [Bai et al. 2024], DitCtrl [Cai et al. 2024], and ReVideo [Mou et al. 2024]. Metrics include masked region preservation, text alignment, and video quality. **Red** stands for the best, **Blue** stands for the second best.

Metrics		Masked Region Preservation					Text Alignment		Video Quality
	Models	PSNR↑	SSIM↑	LPIPS$_{\times 10^2}$↓	MSE$_{\times 10^2}$↓	MAE$_{\times 10^2}$↓	CLIP Sim↑	CLIP Sim (M)↑	FVID↓
Standard	UniEdit	9.96	0.36	56.68	11.08	25.78	8.46	14.23	1.36
	DitCtrl	9.30	0.33	57.42	12.73	27.45	8.52	15.59	0.57
	ReVideo	**15.52**	**0.49**	**27.68**	**3.49**	**11.14**	**9.34**	**20.01**	**0.42**
	Ours	**22.63**	**0.91**	**7.65**	**1.02**	**2.90**	**8.67**	**20.20**	**0.18**
Long	UniEdit	10.37	0.30	54.61	10.25	24.89	10.85	15.42	1.00
	DitCtrl	9.76	0.28	62.49	11.50	26.64	**11.78**	16.52	0.56
	ReVideo	**15.50**	**0.46**	**28.57**	**3.92**	**12.24**	11.22	**20.50**	**0.35**
	Ours	**22.60**	**0.90**	**7.53**	**0.86**	**2.76**	**11.85**	**19.38**	**0.11**

4.4 Human Evaluation

We conducted a user study on video inpainting and editing tasks using standard-length video samples from the *VPBench* inpainting and editing subsets. Thirty participants evaluated 50 randomly selected cases based on background preservation, text alignment, and video quality. As shown in Tab. 4, *VideoPainter* significantly outperformed existing baselines, achieving higher preference rates across all evaluation criteria in both tasks. Detailed experiment settings and results are provided in the Appendix.

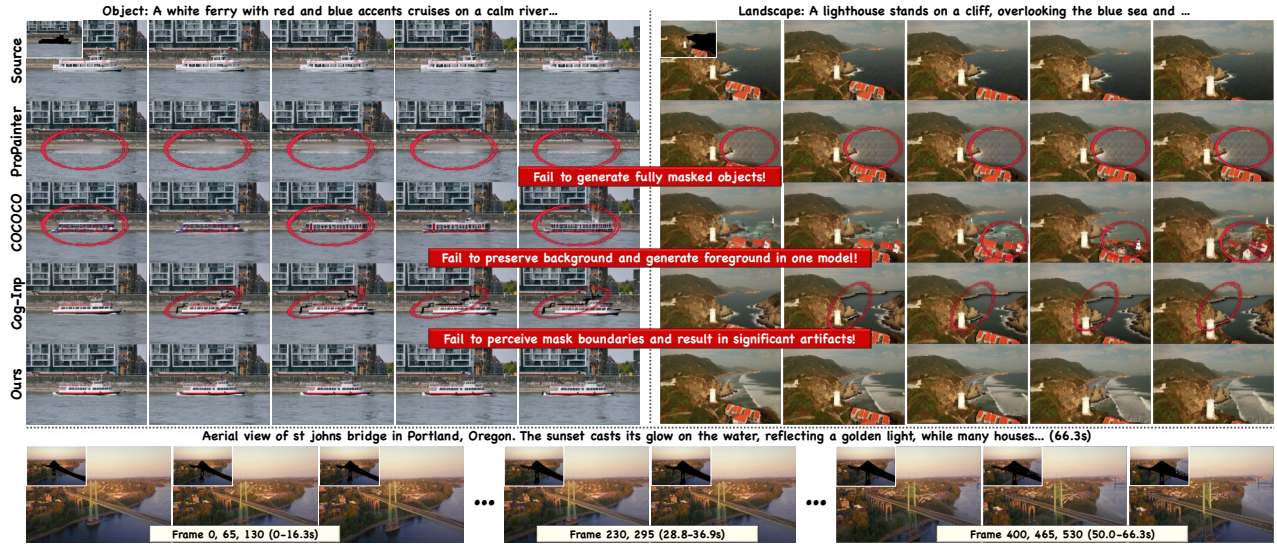

Fig. 5. Comparison of previous inpainting methods and *VideoPainter* on standard and long video inpainting. More visualizations are in the demo video.

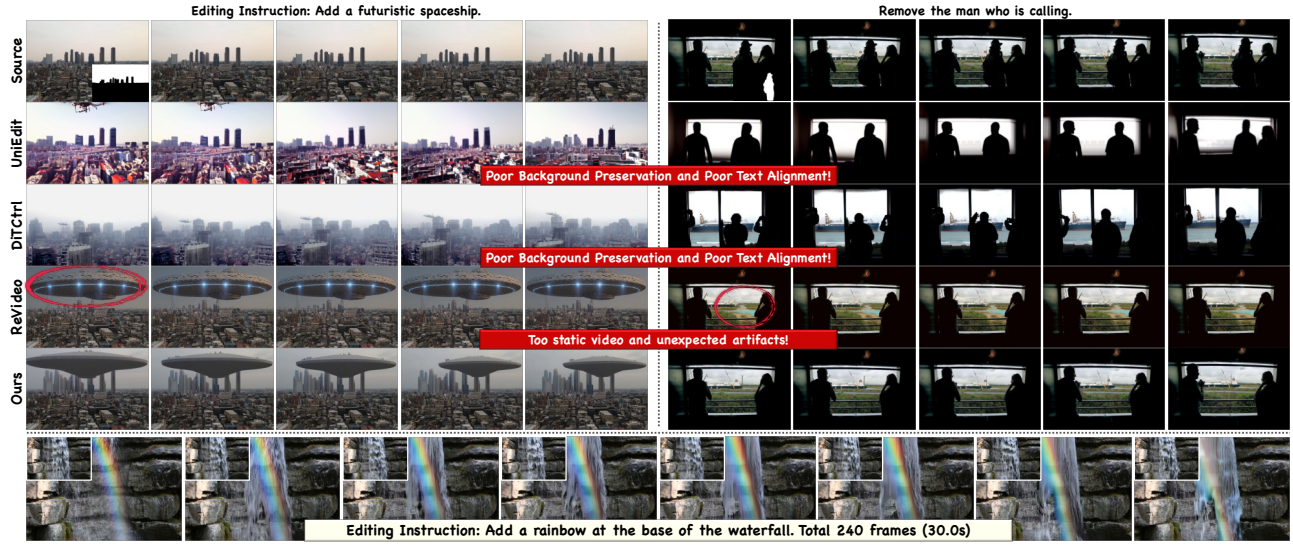

Fig. 6. Comparison of previous editing methods and VideoPainter on standard and long video editing. More visualizations are in the demo video.

4.5 Ablation Analysis

We ablate on *VideoPainter* in Tab .5, including architecture, context encoder size, control strategy, and inpainting region ID resampling.

Based on rows 1 and 5, the dual-branch *VideoPainter* significantly outperforms its single-branch counterpart by explicitly decoupling background preservation from foreground generation, thereby reducing model complexity and avoiding the trade-off between competing objectives in a single branch. Row 2 to row 6 of Tab. 5 demonstrate the rationale behind our key design choices: ❶ utilizing a two-layer structure as an optimal balance between performance and efficiency for the context encoder, ❷ implementing token-selective feature fusion based on segmentation mask information to prevent confusion from indistinguishable foreground-background tokens in the backbone, and ❸ adapting plug-and-play control to different backbones with comparable performance. Furthermore, rows 7 and 8 verify the importance of employing inpainting region ID resampling for long videos, which maintains ID consistency by explicitly resampling inpainted region tokens from previous clips.

4.6 Plug-and-Play Control Ability

Fig. 7 demonstrates the flexible plug-and-play control capabilities of *VideoPainter* in base diffusion transformer selection. We showcase

Table 4. **User Study: User preference ratios comparing *VideoPainter* with video inpainting and editing baselines.** For each sample, participants selected only one model that produced the best results for each criterion. We evaluate performance using the average proportion of being selected as the best response. For video inpainting, we compared *VideoPainter* against ProPainter [Zhou et al. 2023], COCOCO [Zi et al. 2024], and Cog-Inp [Yang et al. 2024]. For video editing, we compared *VideoPainter* against UniEdit [Bai et al. 2024], DitCtrl [Cai et al. 2024], and ReVideo [Mou et al. 2024]. Detailed results are in the appendix.

Task	Video Inpainting			Video Editing		
	Background Preservation	Text Alignment	Video Quality	Background Preservation	Text Alignment	Video Quality
Ours	74.2%	82.5%	87.4%	78.4%	76.1%	81.7%

Table 5. **Ablation Studies on *VPBench*. Single-Branch:** We add input channels to adapt masked video and finetune the backbone. **Layer Configuration (*VideoPainter* (*)):** We vary the context encoder depth from one to four layers. **w/o Selective Token Integration (w/o Select):**: We bypass the token pre-selection step and integrate all context encoder tokens into DiT. **T2V Backbone (*VideoPainter* (T2V)):** We replace the backbone from image-to-video DiTs to text-to-video DiTs. **w/o target region ID resampling (w/o Resample):** We ablate on the target region ID resampling. (L) denotes evaluation on the long video subset. **Red** stands for the best result.

Metrics	Masked Region Preservation					Text Alignment		Video Quality
Models	PSNR↑	SSIM↑	LPIPS$_{\times 10^2}$↓	MSE$_{\times 10^2}$↓	MAE$_{\times 10^2}$↓	CLIP Sim↑	CLIP Sim (M)↑	FVID↓
Single-Branch	20.54	0.79	10.48	0.94	4.16	8.19	19.31	0.22
VideoPainter (1)	21.92	0.81	8.78	0.89	3.26	8.44	20.79	0.17
VideoPainter (4)	22.86	0.85	**6.51**	0.83	2.86	9.12	20.49	0.16
w/o Select	20.94	0.74	7.90	0.95	3.87	8.26	17.84	0.25
VideoPainter (T2V)	23.01	0.87	6.94	0.89	2.65	**9.41**	20.66	0.16
VideoPainter	**23.32**	**0.89**	6.85	**0.82**	**2.62**	8.66	**21.49**	**0.15**
w/o Resample (L)	21.79	0.84	**8.65**	0.81	3.10	11.35	20.68	0.19
VideoPainter (L)	**22.19**	**0.85**	9.14	**0.71**	**2.92**	**11.52**	**21.54**	**0.17**

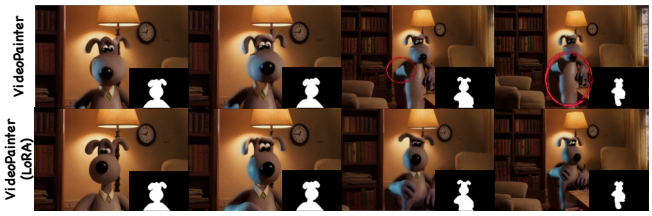

Fig. 7. Integrating *VideoPainter* to Gromit-style LoRA [Cseti 2024].

how *VideoPainter* can be seamlessly integrated with community-developed Gromit-style LoRA. Despite the significant domain gap between anime-style data and our training dataset, *VideoPainter*'s dual-branch architecture ensures its plug-and-play inpainting abilities, enabling users to select the most appropriate base model for specific inpainting requirements and expected results.

5 DISCUSSION

In this paper, we introduce *VideoPainter*, the first dual-branch video inpainting framework with plug-and-play control capabilities. Our approach features three key innovations: (1) a lightweight plug-and-play context encoder compatible with any pre-trained video DiTs, (2) an inpainting region ID resampling technique for maintaining long video ID consistency, and (3) a scalable dataset pipeline that produced *VPData* and *VPBench*, containing over 390K video clips with precise masks and dense captions. *VideoPainter* also shows promise in video editing applications. Extensive experiments demonstrate that *VideoPainter* achieves state-of-the-art performance across 8 metrics in video inpainting and editing, particularly in video quality, mask region preservation, and text coherence.

However, *VideoPainter* still has limitations: (1) Generation quality is limited by the base model, which may struggle with complex physical and motion modeling, and (2) performance is suboptimal with low-quality masks or misaligned video captions.

ACKNOWLEDGMENTS

This work was supported in part by the CUHK Strategic Seed Funding for Collaborative Research Scheme under Grant No. 3136023 and in part by the CUHK Research Matching Scheme under Grant No. 7106937, 8601130, and 8601440.

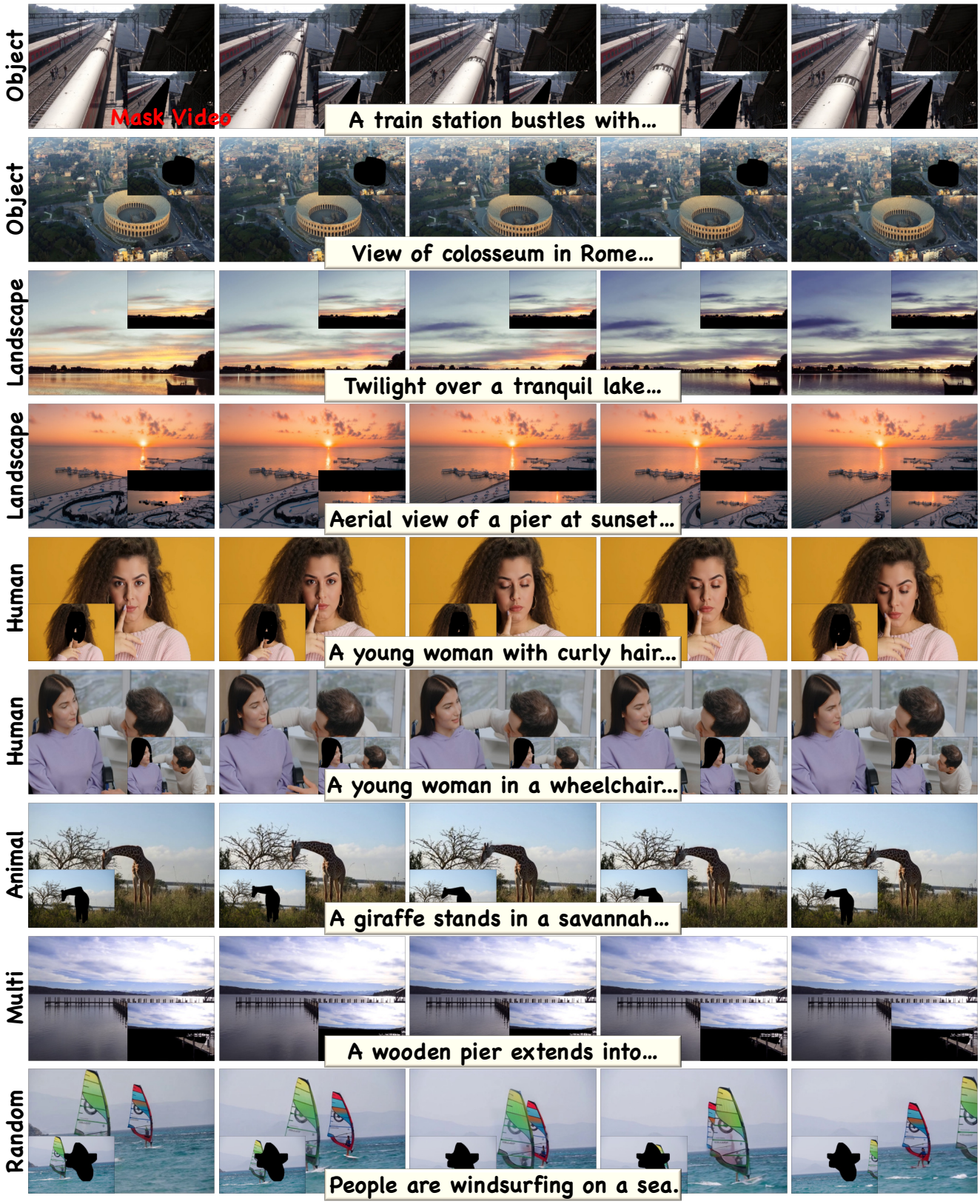

Fig. 8. More video inpainting results.

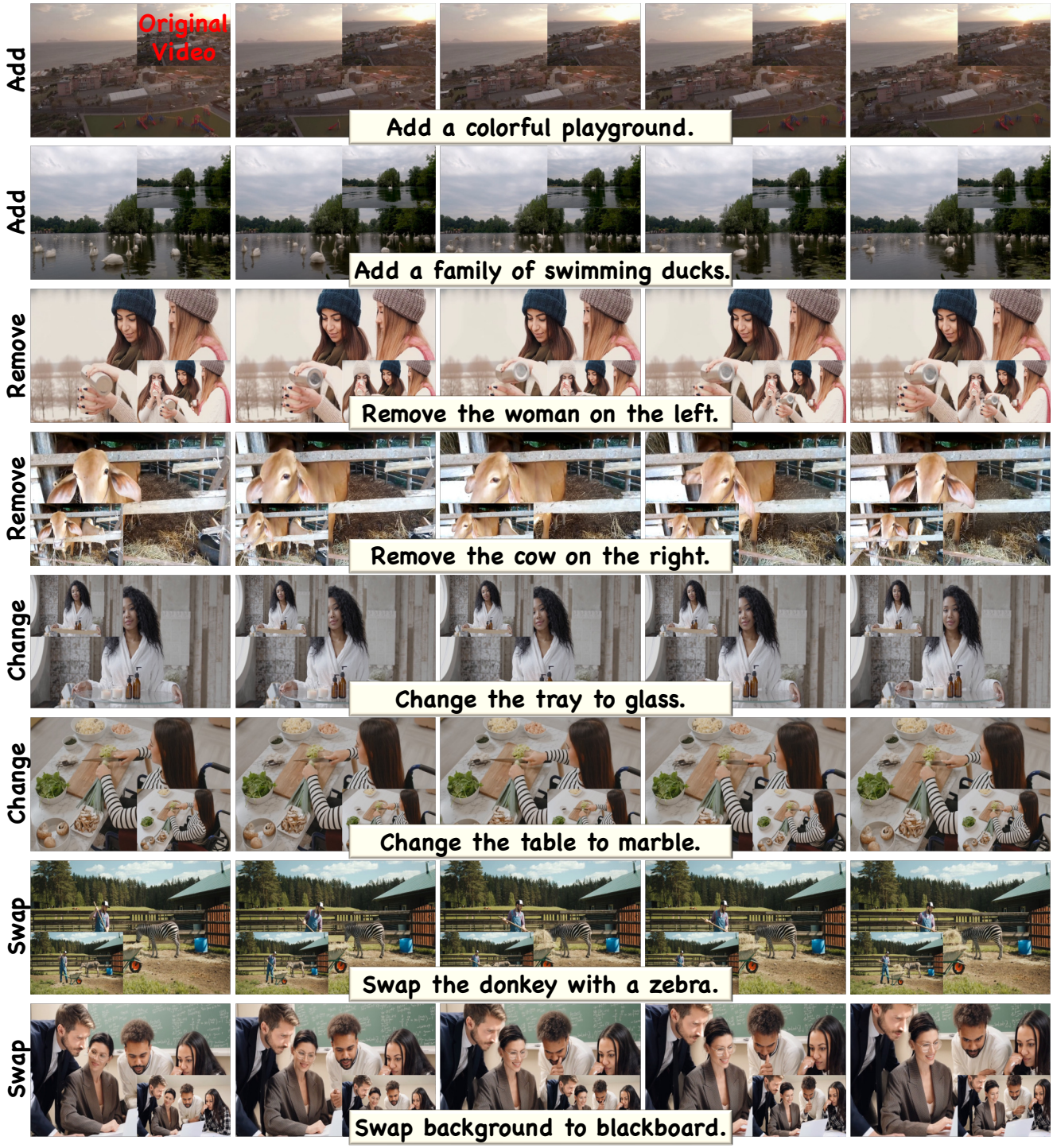

Fig. 9. More video editing results.

REFERENCES

Omri Avrahami, Ohad Fried, and Dani Lischinski. 2023. Blended latent diffusion. *ACM transactions on graphics (TOG)* 42, 4 (2023), 1–11.

Jianhong Bai, Tianyu He, Yuchi Wang, Junliang Guo, Haoji Hu, Zuozhu Liu, and Jiang Bian. 2024. Uniedit: A unified tuning-free framework for video motion and appearance editing. *arXiv preprint arXiv:2402.13185* (2024).

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2, 3 (2023), 8.

Yuxuan Bian, Xuan Ju, Jiangtong Li, Zhijian Xu, Dawei Cheng, and Qiang Xu. 2024. Multi-patch prediction: Adapting llms for time series representation learning. *arXiv preprint arXiv:2402.04852* (2024).

Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. 2024. DiTCtrl: Exploring Attention Control in Multi-Modal Diffusion Transformer for Tuning-Free Multi-Prompt Longer Video Generation. *arXiv preprint arXiv:2412.18597* (2024).

Brandon Castellano. 2024. *PySceneDetect: Intelligent Scene Cut Detection and Video Analysis Tool.* https://github.com/Breakthrough/PySceneDetect

Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. 2019a. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9066–9075.

Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. 2019b. Learnable gated temporal shift module for deep video inpainting. *arXiv preprint arXiv:1907.01131* (2019).

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. arXiv:2310.00426 [cs.CV]

Cseti. 2024. *CogVideoX-LoRA-Wallace_and_Gromit.* Hugging Face. https://huggingface.co/Cseti/CogVideoX-LoRA-Wallace_and_Gromit

Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. 2022. EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. 2023. MOSE: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20224–20234.

Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. 2024. ViViD: Video Virtual Try-on using Diffusion Models. *arXiv preprint arXiv:2405.11794* (2024).

Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. 2020. Flow-edge guided video completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 713–729.

Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, et al. 2024. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*. 1–12.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023).

Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. 2023. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13480–13492.

Yuan-Ting Hu, Heng Wang, Nicolas Ballas, Kristen Grauman, and Alexander G Schwing. 2020. Proposal-based video completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer, 38–54.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. 2024. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976* (2024).

Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. 2019. Deep video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5792–5801.

Kuaishou. 2024. KLING SPARK YOUR IMAGINATION. https://kling.kuaishou.com/.

Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. 2019. Copy-and-paste networks for deep video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4413–4421.

Ang Li, Shanshan Zhao, Xingjun Ma, Mingming Gong, Jianzhong Qi, Rui Zhang, Dacheng Tao, and Ramamohanarao Kotagiri. 2020. Short-term and long-term context aggregation network for video inpainting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 728–743.

Yaowei Li, Yuxuan Bian, Xuan Ju, Zhaoyang Zhang, Ying Shan, and Qiang Xu. 2024. BrushEdit: All-In-One Image Inpainting and Editing. *arXiv preprint arXiv:2412.10316* (2024).

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. 2022. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 17562–17571.

Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. 2021. Decoupled spatial-temporal transformer for video inpainting. *arXiv preprint arXiv:2104.06637* (2021).

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

Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. 2024. ReVideo: Remake a Video with Motion and Content Control. *arXiv preprint arXiv:2405.13865* (2024).

NVIDIA. 2025. NVIDIA Cosmos: Accelerate Physical AI Development with World Foundation Models. https://www.nvidia.com/en-us/ai/cosmos/

OpenAI. 2024. *Hello GPT-4o.* https://openai.com/index/hello-gpt-4o/

OpenAI. 2024. Video generation models as world simulators. https://openai.com/sora/.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 724–732.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. 2024. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720* (2024).

Weize Quan, Jiaxi Chen, Yanli Liu, Dong-Ming Yan, and Peter Wonka. 2024. Deep learning-based image and video inpainting: A survey. *International Journal of Computer Vision* 132, 7 (2024), 2367–2400.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Robin Rombach and Patrick Esser. 2022. Stable Diffusion 2 Inpainting. https://huggingface.co/stabilityai/stable-diffusion-2-inpainting.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. 2024. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, and Dacheng Tao. 2024. Diffusion model-based video editing: A survey. *arXiv preprint arXiv:2407.07111* (2024).

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 402–419.

Pavel Tokmakov, Jie Li, and Adrien Gaidon. 2023. Breaking the "Object" in Video Object Segmentation. In *CVPR*.

Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. 2019. Video inpainting by jointly learning temporal structure and spatial details. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5232–5239.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/d86ea612dec96096c5e0fcc8dd42ab6d-Paper.pdf

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. CogVLM: Visual Expert for Pretrained Language Models. arXiv:2311.03079 [cs.CV]

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. 2024. Videocomposer: Compositional

video synthesis with motion controllability. *Advances in Neural Information Processing Systems* 36 (2024).

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

Wikipedia contributors. 2024a. Mean absolute error — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Mean_absolute_error

Wikipedia contributors. 2024b. Mean squared error — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Mean_squared_error

Wikipedia contributors. 2024c. Peak signal-to-noise ratio — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Peak_signal-to-noise_ratio&oldid=1210897995 [Online; accessed 4-March-2024].

Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. 2021. GODIVA: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806* (2021).

Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. 2018. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 585–601.

Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. 2019. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3723–3732.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072* (2024).

Yanhong Zeng, Jianlong Fu, and Hongyang Chao. 2020. Learning joint spatial-temporal transformations for video inpainting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*. Springer, 528–543.

Kaidong Zhang, Jingjing Fu, and Dong Liu. 2022a. Flow-guided transformer for video inpainting. In *European Conference on Computer Vision*. Springer, 74–90.

Kaidong Zhang, Jingjing Fu, and Dong Liu. 2022b. Flow-guided transformer for video inpainting. In *European Conference on Computer Vision*. Springer, 74–90.

Kaidong Zhang, Jingjing Fu, and Dong Liu. 2022c. Inertia-guided flow completion and style fusion for video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5982–5991.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 586–595.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2024a. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1724–1732.

Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. 2024b. AVID: Any-Length Video Inpainting with Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7162–7172.

Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. 2023. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10477–10486.

Bojia Zi, Shihao Zhao, Xianbiao Qi, Jianan Wang, Yukai Shi, Qianyu Chen, Bin Liang, Kam-Fai Wong, and Lei Zhang. 2024. CoCoCo: Improving Text-Guided Video Inpainting for Better Consistency, Controllability and Compatibility. *arXiv preprint arXiv:2403.12035* (2024).

Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. 2021. Progressive temporal feature alignment network for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16448–16457.