O-DisCo-Edit: Object Distortion Control for Unified Realistic Video Editing

Yuqing Chen[1,3*], **Junjie Wang**[1†], **Lin Liu**[2‡], **Ruihang Chu**[1],
Xiaopeng Zhang[2], **Qi Tian**[2], **Yujiu Yang**[1]
[1]Tsinghua University
[2]Huawei Inc.
[3]Pengcheng National Laboratory
chenyuqi24@mails.tsinghua.edu.cn, wangjunjie@sz.tsinghua.edu.cn, ll0825@mail.ustc.edu.cn

Abstract

Diffusion models have recently advanced video editing, yet controllable editing remains challenging due to the need for precise manipulation of diverse object properties. Current methods require different control signal for diverse editing tasks, which complicates model design and demands significant training resources. To address this, we propose O-DisCo-Edit, a unified framework that incorporates a novel object distortion control (O-DisCo). This signal, based on random and adaptive noise, flexibly encapsulates a wide range of editing cues within a single representation. Paired with a "copy-form" preservation module for preserving non-edited regions, O-DisCo-Edit enables efficient, high-fidelity editing through an effective training paradigm. Extensive experiments and comprehensive human evaluations consistently demonstrate that O-DisCo-Edit surpasses both specialized and multitask state-of-the-art methods across various video editing tasks.

Demo — https://cyqii.github.io/O-DisCo-Edit.github.io/
Extended version — https://arxiv.org/abs/2509.01596

1 Introduction

Recent years have witnessed remarkable advancements in diffusion-based video generation (Yang et al. 2024b; Wan et al. 2025; Hong et al. 2022; HaCohen et al. 2024). Beyond pure generation, video editing has emerged as a crucial extension, which enables modifications to reference videos based on user instructions. Specifically, effective video editing necessitates precise control over the content within edited regions, while flawlessly preserving unedited areas.

For controllable video editing, single-task editing models incorporate additional control signals such as 2D bounding boxes (Tu et al. 2025; Yang et al. 2024a; Li et al. 2025a), masks (Zhao, Ma, and Zhou 2025; Yariv et al. 2025; Huang et al. 2025), optical flow (Yin et al. 2023; Wang et al. 2025a; Liu et al. 2024a), and tracking points (Gu et al. 2025; Wu et al. 2024; Wang et al. 2025b) to improve control precision.

However, as shown in Fig. 1, conditions like bounding boxes and masks provide limited information, thus hindering fine-grained control for complex editing scenarios. Furthermore, video datasets with optical flow and tracking are scarce, and their extraction is often complex and prone to inaccuracies. These two issues make precise and intricate controllable editing difficult.

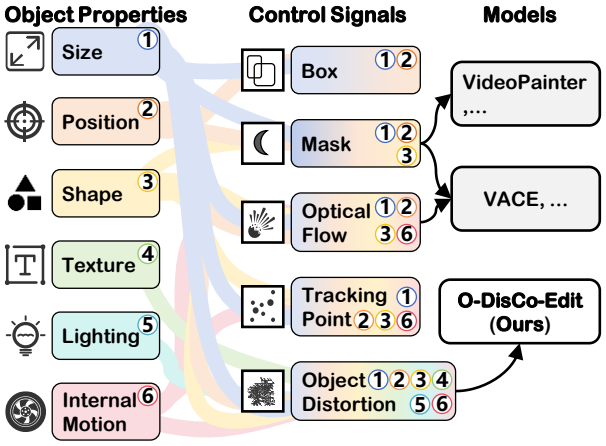

Figure 1: Comparisons of different object properties, control signals, and models.

Model	Dataset	Trainable Module	Steps	GPUs
VACE	Mutil-Task	8 Blocks	200K	128 A100
Senorita	Mutil-Task	102 Blocks*	4 epoch	\
VideoPainter	390.3k	2 Blocks, 1 LoRA	82K	64 V100
Ours	180k	Two LoRAs	7.55K	8 A800

Table 1: Comparison of training configurations for different models. * indicates that the majority of the module is used for training. "Block" refers to a DiT block.

Single-task editing models, as discussed above, are no longer sufficient to meet user diverse demands. Consequently, unified multi-task video editing approaches (Liang et al. 2025; Li et al. 2025b; Zhang et al. 2025c; Ye et al. 2025; Jiang et al. 2025) have emerged, which can accomplish diverse editing tasks by introducing various signals. However, they typically demand complex training pipelines. This complexity necessitates the construction of specialized multi-task datasets and the design of task-specific modules

*This work was done during an internship at Huawei Inc.
†Corresponding author.
‡Corresponding author and project leader.

(e.g., multiple DiT blocks), resulting in a large number of trainable parameters, as shown in Tab. 1. Furthermore, it requires integrating various conditions across numerous training stages, often demanding tens of thousands of steps.

Despite their design incorporating diverse signals, most of multi-task video editing model are inflexible during inference, as they can generally process only one control condition at a time. This prevents the model from leveraging complementary cues from multiple signals, thereby hindering the flexible transition between fine-grained and coarse-grained editing for the same task.

To address these challenges, we propose a novel unified control signal: the **o**bject **dis**tortion **co**ntrol (O-DisCo). This signal is generated by applying appropriate noise to the edited objects, effectively acting as a distortion signal for the reference video. As illustrated in Fig. 1, all other control signals can similarly be viewed as specific types of reference video distortion. Therefore, by controlling the noise, O-DisCo inherently unifies all these signals into a single representation. For training, randomness is introduced into O-DisCo. This significantly simplifies training dataset construction and model design, saving substantial training resources, as shown in Tab. 1. For inference, by adaptively manipulating the intensity and scope of O-DisCo's noise, our model can perform a wide range of tasks. While the above design primarily focuses on the edited regions, a "copy-form" preservation module is further designed to address the preservation of non-edited areas. Encapsulating these capabilities, we propose O-DisCo-Edit, a unified framework for versatile video editing.

Comprehensive experiments confirm O-DisCo-Edit's effectiveness and versatility across diverse tasks, including object removal, outpainting, color change and transfers of motion, lighting and style. Specifically, O-DisCo-Edit consistently surpasses the state-of-the-art (SOTA) multi-task editing model VACE (Jiang et al. 2025), on the majority of tasks. Notably, for the object removal on the OmnimatteRF (Lin et al. 2023) benchmark, our method also demonstrates superior performance compared to the specialized SOTA removal approach, MiniMax-Remover (Zi et al. 2025a).

Overall, our contributions are summarized as:

- A novel unified control signal, **o**bject **dis**tortion **co**ntrol (O-DisCo) is proposed to substantially reduce training resource demands and enable flexible, precise multi-task video editing from coarse to fine granularity.
- We propose a "copy-form" preservation module for non-edited region preservation, which enhances the model's ability to maintain unedited areas.
- Our proposed O-DisCo-Edit, achieving new SOTA performance across diverse tasks, offers a novel perspective for developing unified video editing frameworks.

2 Related Work

Single-Task Video Editing and Control Signals. Video editing tasks frequently require additional control signals (e.g., masks, poses, optical flows, tracking points) to modify reference video attributes. VideoAnydoor (Tu et al. 2025) introduces masks and tracking points for object insertion, while DiffuEraser (Gu et al. 2025) leverage masks for object removal. Follow-your-Canvas (Chen et al. 2024) employs 2D boxes for outpainting. Bai et al. (2025) and Liu et al. (2024b) use camera trajectories for camera control.

Multi-Task Video Editing. Growing demands for creative versatility have driven the development of multi-task video editing. VACE (Jiang et al. 2025) integrates sophisticated signals like optical flow and masks with a context embedder and adapter to perform tasks such as swap, animation, and outpainting. Similarly, Senorita (Zi et al. 2025b) utilizes masks, canny edges, and other cues, coupled with four specialized expert models, to achieve tasks like addition, removal, and swap. Therefore, multi-task editing often demands complex training pipelines with diverse signals, specialized modules, and multi-stage training (Ye et al. 2025; Liang et al. 2025; Jiang et al. 2025; Zi et al. 2025b). In contrast, our proposed unified O-DisCo signal enables multi-task completion with significantly fewer training resources.

Adaptive Inference. Current video editing models (Jiang et al. 2025; Zi et al. 2025b; Liang et al. 2025; Ye et al. 2025; Tu et al. 2025) typically rely on a single control signal during inference, which limits their adaptability for multi-grained editing. Adaptive inference, conversely, allows models to adjust outputs dynamically based on varying reference videos, images, or prompts. This is common in training-free models (Pan et al. 2025; Chen et al. 2025a; Zhang et al. 2025b), LMP (Chen et al. 2025a) applies attention values as a loss to optimize hidden states for appearance similarity. Likewise, Zhang et al. (2025b) defines motion consistency loss for gradient descent on noisy latent vectors to achieve motion transfer. Building on this principle, O-DisCo-Edit also leverages adaptive inference through the proposed O-DisCo, which dynamically adjusts injected noise based on reference videos and images for multi-grained control.

3 Methodology

As shown in Fig. 2, our approach introduces a first-frame-guided video editing model, building upon the CogVideoX-I2V (Yang et al. 2024b). The object distortion control (O-DisCo), derived via the distorter, is fed into the VAE and conditional DiT for precise control over edited regions. Concurrently, a "copy-form" preservation (CFP) module processes the reference image and video, which then provide its latent output to the denoising DiT for robust preservation of non-edited areas. Additionally, an identity preservation (IDP) module is proposed to enhance ID fidelity within edited regions. Subsequent sections detail O-DisCo's construction and the design of the CFP and IDP modules.

3.1 Random Object Distortion Control

During the training phase, we apply a random distoter to generate random object distortion control (R-O-DisCo). As shown in the top-left part of Fig. 2, we intentionally distort the colors of the reference video $V_{\text{ref}} \in \mathbb{R}^{F\times H\times W\times 3}$ to prevent the model from simply replicating original color information, where F is the frame number, H and W are the height and width of the reference video. This involves applying random arithmetic operations to each RGB channel,

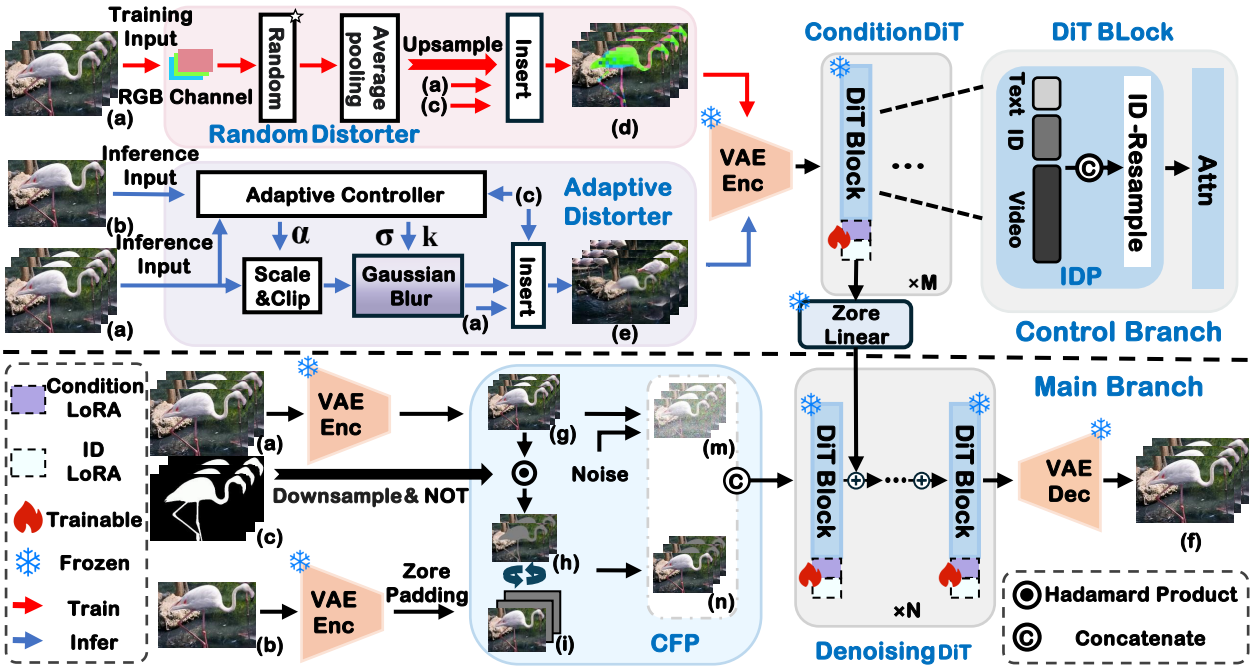

Figure 2: The framework of the proposed O-DisCo-Edit. (a) Reference video. (b) Reference image (first frame during training, edited image during inference). (c) Masks. (d) R-O-DisCo. (e) A-O-DisCo. (f) Generated video. (g) Latent of reference video. (h) Latent of the preserved region. (i) Image latent with zero-padding. (m) Noisy Latent. (n) Image Latent with the latent of preserved region. α is the contrast, σ represents the intensity of the added noise, and k is the size of the gaussian blur kernel.

and the resulted color-distorted video $V_{\text{cd}} \in \mathbb{R}^{F\times H\times W\times 3}$ is formulated as:

$$V_{\text{cd}}[:,:,:,i] = \text{clip}(V_{\text{ref}}[:,:,:,i] \star a_i),\ i \in \{0,1,2\}, \quad (1)$$

where a_i is random real number. Moreover, $\star$ denotes a randomly chosen arithmetic operation (addition, subtraction, multiplication, or division). The clip operation constrains output values to the range $[0, 255]$.

After that, V_{cd} is downsampled via average pooling and then upsampled using nearest-neighbor interpolation, both by a randomly sampled factor of $L \in \mathbb{Z}$. This operation intentionally disrupts fine-grained structural details, producing a mosaic-like effect video $V_{\text{cdm}} \in \mathbb{R}^{F\times H\times W\times 3}$. Consequently, the model is compelled to primarily learn video generation guided by the first frame's appearance, rather than relying on the precise visual information of V_{cd}.

Finally, We utilize masks $M \in \mathbb{R}^{F\times H\times W\times 1}$ to insert the edited object from V_{cdm} into V_{ref}, which produces the R-O-DisCo V_{RODC} ((d) in Fig. 2) during the training stage:

$$V_{\text{RODC}} = V_{\text{cdm}} \odot M + V_{\text{ref}} \odot (\mathbf{1} - M), \quad (2)$$

where $\odot$ represents the Hadamard product.

Overall, during the training phase, we enhance the model's robustness and task adaptability by increasing the randomness of O-DisCo. Detailed parameter ranges are in Appendix A of our extended version (Chen et al. 2025b).

3.2 Adaptive Object Distortion Control

During inference phase, our model adapts to specific tasks or instructions via adaptive object distortion control (A-O-DisCo) (highlighted by (e) in Fig. 2), which is implemented by an adaptive distorter. It is achieved through contrast modification (scaling and clipping) and dynamic noise injection within the editable regions. The process is formally represented by the following equations:

$$\begin{aligned} V_{\text{c}}(f,x,y) &= \text{clip}\big(\alpha \cdot V_{\text{ref}}(f,x,y)\big), \\ V_{\text{cn}}(f,x,y) &= \sum_{i=-b}^{b}\sum_{j=-b}^{b} V_{\text{c}}(f,x+i,y+j)\cdot G_{\text{norm}}(i,j;\sigma), \\ V_{\text{AODC}} &= V_{\text{cn}} \odot M + V_{\text{ref}} \odot (\mathbf{1} - M), \end{aligned} \quad (3)$$

where $V_{\text{ref}}(f,x,y)$ denotes the pixel value at coordinates (x,y) in the f-th frame of the reference video. $V_{\text{c}}(x,y)$ and $V_{\text{cn}}(x,y)$ represent the V_{ref} after scale&clip and gaussian blur, respectively. $G_{\text{norm}}(i,j;\sigma)$ is the normalized gaussian blur kernel. α represents the contrast, σ is the noise intensity, and $k = 2b+1$ is the gaussian kernel size.

The adaptive controller determines suitable values for α, σ, and k by calculating two similarities: $\mathbf{Sim}_{\text{i}}$, the edited region's edge map similarity between the reference image and the reference video's first frame; $\mathbf{Sim}_{\text{v}}$, the intra-frame similarity within the reference video's edited region edge map. Empirically, fitting these three parameters using a quadratic polynomial of two similarity yields superior results (see specific formulas in Appendix A of our extended version (Chen et al. 2025b)). Finally, the A-O-DisCo V_{AODC} obtained for each task is shown in the last line of Tab. 2. Specifically, during inference for object removal and outpainting (R&O), we set V_{AODC} to zero, ensuring that no additional information

Condition	R&O	Style Tranfer	Other Tasks
z_{images}	$[z^{\text{i}}_{\text{ref}}, z^{\text{v}'}_{\text{p}}]$	$[z^{\text{i}}_{\text{ref}}, \mathbf{0}]$	$[z^{\text{i}}_{\text{ref}}, z^{\text{v}'}_{\text{p}}]$
V_{AODC}	$\mathbf{0}$	$V_{\text{cn}} \odot M$ $+V_{\text{ref}}(1-M)$	$V_{\text{cn}} \odot M$ $+V_{\text{ref}}(1-M)$

Table 2: Adaptive inference conditions for different tasks. R&O in the first line means object removal and outpainting.

is introduced. This allows the model to generate the video based on other condition, thereby reducing artifacts.

3.3 "Copy-Form" Preservation Module

Many video editing methods (Jiang et al. 2025; Bian et al. 2025; Liang et al. 2025) typically integrate non-edited regions with the control signals via extra branch. However, such integration often leads to mutual interference between the preserved regions and control signals, thereby limiting editing flexibility.

Instead, we propose the "copy-form" preservation (CFP) module illustrated Fig. 2, which enhances editing flexibility by integrating the non-edited regions directly into the main branch of the network. Detailedly, CFP replaces conventional zero-padding (denoted as (i) in Fig. 2) with the latent of the preserved region $z^{\text{v}'}_{\text{p}}$ (marked as (h) in Fig. 2), to obtain z_{images} (denoted as (n) in Fig. 2). This process is expressed as:

$$\begin{aligned} z^{\text{v}'}_{\text{p}} &= z^{\text{v}}_{\text{ref}} \odot (\mathbf{1} - z_{\text{mask}})[1:], \\ z_{\text{images}} &= [z^{\text{i}}_{\text{ref}}, z^{\text{v}'}_{\text{p}}], \end{aligned} \tag{4}$$

where $z^{\text{v}}_{\text{ref}}$ denotes the latent of the reference video, z_{mask} represents the downsampled binary mask (with the same shape as $z^{\text{v}}_{\text{ref}}$), $z^{\text{i}}_{\text{ref}}$ is reference image latent. $[1:]$ corresponds to a slicing operation, and $[,]$ signifies tensor concatenation. The z_{images} for each task are shown in Tab. 2. Notably, the CFP module achieves preservation of non-edited regions with an effect similar to "first-frame copying".

3.4 Identity Preservation Module

To mitigate object appearance changes during complex motion, we designed the identity preservation (IDP) module, illustrated in the upper right corner of Fig. 2. Specifically, we extract position-agnostic tokens (ID tokens) from the reference image's edited regions and concatenate them with text tokens. Akin to text tokens, ID tokens act as a global guide, ensuring the model leverages ID information throughout video generation. To further enhance ID consistency, ID-Resample extracts the key (K) and value (V) vectors from the edited regions of the generated video. These are concatenated with original generated video's the K and V vectors , and the process compels the model to reinforce ID consistency within the edited regions.

4 Experiments

4.1 Experimental Setup

Implementation Details. Training dataset: we utilize approximately 180k video-mask pairs from the Senorita-2M grounding dataset (Zi et al. 2025b). All video-mask pairs are center-cropped and resized to a 720×480 resolution with a length of 49 frames. Moreover, prompts for masked regions are generated via Qwen2.5-VL-7B (Bai et al. 2023). Two-stage training: Our model builds on the frozen pre-trained weights of Diffusion as Shader (Gu et al. 2025). Firstly, condition LoRA is trained with the random distorter and CFP module (2400 steps); Secondly, the IDP module's dedicated ID LoRA is trained (5150 steps). All stages employ AdamW optimization (learning rate 1×10^{-4}) on 8 A800 GPUs with gradient accumulation for a batch size of 32.

Baseline Methods. For the majority of tasks, we select the SOTA unified video editing methods VACE (Jiang et al. 2025), VideoPainter (Bian et al. 2025), and Senorita (Zi et al. 2025b) as our primary baselines. Additionally, for the object removal, we include DiffuEraser (Li et al. 2025c), MiniMax-Remover (MiniMax) (Zi et al. 2025a), and Propainter (Zhou et al. 2023) as extra baselines. For the style transfer, Senorita (Zi et al. 2025b) is chosen.

Benchmarks. We curated a benchmark from DIVAS (Pont-Tuset et al. 2017) and VPData (Bian et al. 2025), specifically targeting challenging scenarios with internal motion, lighting variations, and complex object movements. For prompts, Senorita utilized the dedicated prompt, as required by its inference process, while others used same prompts from Qwen2.5VL (Bai et al. 2023). Additionally, the reference image was the first frame edited using either HiDream-E1 (Cai et al. 2025) or commercial models[1]. Subsequently, the edited frame served as input for O-DisCo-Edit and multi-task baselines. For fair comparison with Senorita, only the first 33 frames of generated videos are evaluated. In parallel, OmnimatteRF (Lin et al. 2023) is selected as a benchmark for the object removal task.

Metrics. The evaluation includes automatic scoring and a manual user study. Automatic scoring metrics: (1) Non-Edited Region Preservation: Fidelity in unedited regions is assessed using PSNR (PSNR_P) and SSIM (SSIM_P). (2) Alignment: CLIP Similarity (Wu et al. 2021) (CLIP-T) measures semantic consistency between the generated video and its caption. Appearance consistency (Zhang et al. 2025a) (CLIP-I_E) between the output video and the reference image is calculated within the edited regions. (3) Video Generation Quality: Overall video quality is assessed via FVD (Unterthiner et al. 2018), ArtFID (Wright and Ommer 2022), PSNR, SSIM, and temporal consistency (TC) (Zhang et al. 2025a; Chen et al. 2023). (4) Normalized Average Score: This score is obtained by following Huang et al. (2024) using Min-Max Normalization, with all metrics (excluding CLIP-T) weighted equally. Specifically, for the style transfer task, CFSD (Chung, Hyun, and Heo 2024) is applied to evaluate the preservation of the reference video's content. Meanwhile, for the object removal task, we measure removal ability by calculating the SSIM (SSIM_E) and PSNR (PSNR_E) between the edited regions of the output video and the corresponding background video. Manual Assessment: The mean opinion score is adopted, focusing on editing completeness

[1]https://jimeng.jianying.com/

Metrics		Video Quality				Removal Capability		Normalized	User Stuty	
Task	Method	TC↑	FVD↓	PSNR↑	SSIM↑	$SSIM_E$ ↑	$PSNR_E$ ↑	Avg. Score↑	EC↑	VQ↑
(a.1) Object Removal (49)	DiffuEraser	0.9964	422.7	27.89	0.9207	0.9713	34.09	0.2682	3.122	2.867
	MiniMax	0.9973	373.2	27.15	0.8732	**0.9737**	34.87	0.4816	3.567	3.333
	propainter	0.9971	410.3	**28.30**	**0.9224**	0.9715	34.12	0.4844	3.044	2.822
	O-DisCo-Edit	**0.9974**	**300.3**	28.05	0.8751	0.9730	**35.43**	**0.7553**	**3.967**	**3.689**
(a.2) Object Removal (33)	VACE 1.3B	0.9934	1376	23.46	0.8508	0.9551	26.17	0.4233	2.011	1.911
	VACE 14B	0.9896	2085	22.29	0.8372	0.9439	24.28	0.1316	1.578	1.444
	Senorita	0.9962	662.2	26.24	0.8387	0.9681	32.77	0.7058	3.311	3.156
	VideoPainter	0.9871	2403	21.28	0.8303	0.9452	23.42	0.0072	1.744	1.578
	O-DisCo-Edit	**0.9969**	**360.1**	**28.28**	**0.8719**	**0.9740**	**36.05**	**1.000**	**3.956**	**3.689**

Metrics		Video Quality			Alignment	Preservation		Normalized	User Stuty	
Task	Method	TC↑	FVD↓	PSNR↑	CLIP-T↑	$PSNR_P$ ↑	$SSIM_P$ ↑	Avg. Score↑	EC↑	VQ↑
(b) Outpainting	VACE 1.3B	0.9976	88.03	25.11	11.92	31.62	0.9383	0.6801	4.244	3.933
	VACE 14B	0.9977	88.75	23.52	12.03	30.08	0.9381	0.4972	4.178	3.911
	Senorita	**0.9978**	177.1	25.20	\	30.90	0.9262	0.4784	2.889	2.600
	VideoPainter	0.9958	325.4	24.54	**12.95**	31.81	0.9442	0.3384	2.156	1.978
	O-DisCo-Edit	**0.9978**	**77.03**	**26.43**	12.18	**33.87**	**0.9466**	**1.000**	**4.289**	**4.067**

Metrics		Video Quality		Alignment		Preservation		Normalized	User Stuty	
Task	Method	TC↑	ArtFID↓	CLIP-T↑	$CLIP\text{-}I_E$ ↑	$PSNR_P$ ↑	$SSIM_P$ ↑	Avg. Score↑	EC↑	VQ↑
(c) Object Internal Motion Transfer	VACE 1.3B	**0.9946**	7.329	18.75	93.73	**36.41**	**0.9586**	0.8515	3.011	2.533
	VACE 14B	0.9937	7.025	18.96	93.39	34.47	0.9582	0.7641	3.122	2.800
	Senorita	0.9940	**6.628**	\	93.94	29.45	0.8477	0.5229	3.022	2.333
	VideoPainter	0.9908	8.201	**19.47**	91.57	36.32	0.9410	0.3657	2.300	1.822
	O-DisCo-Edit	0.9927	6.712	18.82	**94.64**	35.88	0.9530	**0.8639**	**4.178**	**3.756**
(d) Lighting Transfer	VACE 1.3B	**0.9964**	**5.991**	20.26	95.30	31.46	0.9292	0.7700	3.067	2.644
	VACE 14B	0.9958	6.187	20.35	95.32	30.69	0.9261	0.5489	3.411	3.067
	Senorita	**0.9964**	6.478	\	**96.15**	28.71	0.9011	0.4000	3.033	2.400
	VideoPainter	0.9951	6.378	**21.92**	94.76	32.93	**0.9325**	0.4160	2.911	2.489
	O-DisCo-Edit	0.9956	6.043	20.86	96.05	**33.54**	0.9285	**0.8157**	**3.978**	**3.689**
(e) Color Change	VACE 1.3B	0.9955	8.150	11.89	97.16	30.55	0.9056	0.7838	3.633	3.244
	VACE 14B	0.9954	8.485	11.63	96.55	29.76	0.9036	0.5703	3.456	3.089
	Senorita	**0.9959**	**8.002**	\	**97.67**	27.52	0.8724	0.6000	**4.033**	**3.711**
	VideoPainter	0.9943	9.388	**12.89**	96.41	30.99	**0.9136**	0.3996	3.633	3.267
	O-DisCo-Edit	0.9955	8.008	11.94	97.49	**31.00**	0.9049	**0.8787**	3.944	3.689
Task	Method	TC↑	FVD↓	CLIP-T↑	$CLIP\text{-}I_E$ ↑	$PSNR_P$ ↑	$SSIM_P$ ↑	Avg. Score↑	EC↑	VQ↑
(f) Swap	VACE 1.3B	0.9843	688.2	14.73	91.28	26.36	0.8062	**0.7068**	3.467	2.956
	VACE 14B	0.9841	**642.3**	14.83	91.03	25.76	0.8041	0.6959	3.556	3.089
	Senorita	**0.9845**	803.7	\	**92.76**	23.66	0.7436	0.4000	3.456	2.956
	VideoPainter	0.9815	731.2	**15.91**	90.05	**27.51**	**0.8295**	0.4899	2.967	2.178
	O-DisCo-Edit	0.9839	711.8	15.27	91.84	26.25	0.8098	0.6950	**4.033**	**3.689**
(g) Addition	VACE 1.3B	0.9862	512.7	21.04	92.92	27.47	0.8094	0.3419	3.370	2.489
	VACE 14B	0.9873	398.9	**21.41**	92.58	26.86	0.8082	0.3621	3.200	2.578
	Senorita	**0.9891**	**316.8**	\	**95.39**	27.89	0.7934	**0.7375**	3.496	3.089
	VideoPainter	0.9836	560.9	20.77	93.64	**28.36**	**0.8246**	0.4754	3.104	2.911
	O-DisCo-Edit	0.9871	448.3	21.03	95.37	28.03	0.8048	0.6470	**4.037**	**3.822**
Task	Method	TC↑	ArtFID↓	CLIP-T↑	$CLIP\text{-}I_E$ ↑	$PSNR_P$ ↑	CFSD↓	Avg. Score↑	EC↑	VQ↑
(h) Style Transfer	Senorita	**0.9960**	7.979	\	**94.22**	\	**0.0933**	\	2.989	2.578
	O-DisCo-Edit	0.9954	**7.292**	\	93.72	\	0.2055	\	**4.322**	**4.156**

Table 3: Comparison of different models on various tasks using our benchmark (OmnimatteRF benchmark for object removal). The evaluation includes automatic scoring and a manual user study. The best results are in bold, while the second best are underlined. "Preservation" means non-edited region preservation, "TC" denotes temporal consistency, "EC" represents editing completeness, and "VQ" stands for visual quality.

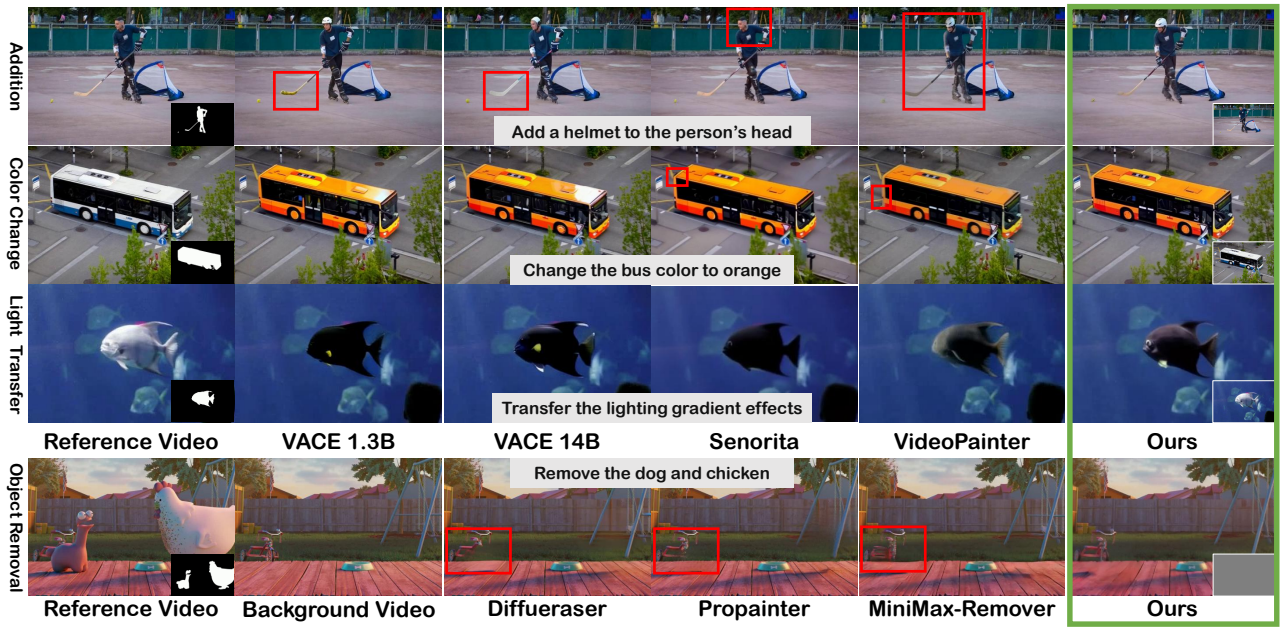

Figure 3: Comparison of our O-DisCo-Edit and baselines for addition, color change, lighting transfer, and object removal. The bottom right of the reference video displays the input masks utilized by all models, while the corresponding position in our results highlights the A-O-DisCo.

(EC) and video quality (VQ). Anonymized generated data is randomly distributed to participants for 1-5 scale scoring.

More details about benchmarks and metrics are in the Appendix B of our extended version (Chen et al. 2025b).

4.2 Comparison with State-of-the-Arts

We conduct comprehensive comparisons between O-DisCo-Edit and the baselines across object removal, outpainting, lighting and object internal motion transfer, color change, swap, addition, and style transfer. Our method demonstrates superior performance in all these tasks.

Object Removal. Quantitatively, our method obtains optimal results in both the Removal (49) (49-frame videos) and Removal (33) (33-frame videos) (Tab. 3 (a)). As shown in Fig. 3, O-DisCo-Edit successfully avoids the background damage seen in Propainter and DiffuEraser, as well as the bicycle overlap present in MiniMax-Remover. In comparison with multi-task models Fig. 4, baselines exhibit prominent artifacts, which indicate unsuccessful removal.

Outpainting. For evaluation, we outpaint videos from 280×520 to 480×720. As demonstrated in Tab. 3 (b), O-DisCo-Edit establishes new SOTA records across all metrics. VACE 1.3B generates grainy textures in edited areas shown in Fig. 5. In contrast, O-DisCo-Edit creates exceptionally well-blended, natural, and continuous results.

Lighting and Object Internal Motion Transfer. O-DisCo-Edit excels at fine-grained editing, including lighting and object internal motion transfer that remain challenging for existing models. As Tab. 3 (c)&(d) illustrate, our approach reaches the best results on both tasks. Correspondingly, as shown in Figs. 3 and 4, VACE 1.3B, VideoPainter, and Senorita fail to handle the subtle changes, while VACE 14B produces results inconsistent with the reference video.

Color Change. O-DisCo-Edit is capable of changing color while preserving intrinsic characteristics. In quantitative analysis, our approach achieves the highest average score as shown in Tab. 3 (e). Qualitatively, a comparison in Fig. 3 reveals that VACE produces irregular color gradients, while Senorita and VideoPainter generate subtle artifacts. Conversely, our approach avoids these issues and yields superior color change results. Notably, Senorita's top score in user studies comes from its first-frame propagation strategy. This strategy creates high visual consistency but performs poorly at preserving non-edited regions.

Swap. Quantitative evaluation in Tab. 3 (f) shows O-DisCo-Edit's performance is second to VACE, yet we achieve a higher CLIP-I_E. As shown in Fig. 4, VACE 14B struggles with ID consistency. Meanwhile, VACE 1.3B and VideoPainter overfit masks boundaries, generating anatomically incorrect outputs (e.g., polar bears with three ears). Furthermore, Senorita, VACE 14B, and VACE 1.3B exhibit motion inconsistencies (red box). Conversely, our method exhibits superior visual results, as evidenced by user study.

Addition. O-DisCo-Edit enables adding new objects to existing moving objects in a video. As shown in Tab. 3 (g), O-DisCo-Edit rank second only to Senorita. However, as Fig. 3 illustrated, Senorita fail to complete the addition task, with its high metrics solely due to "copying" the original video. Therefore, our method attains the most preferred additions results in user study.

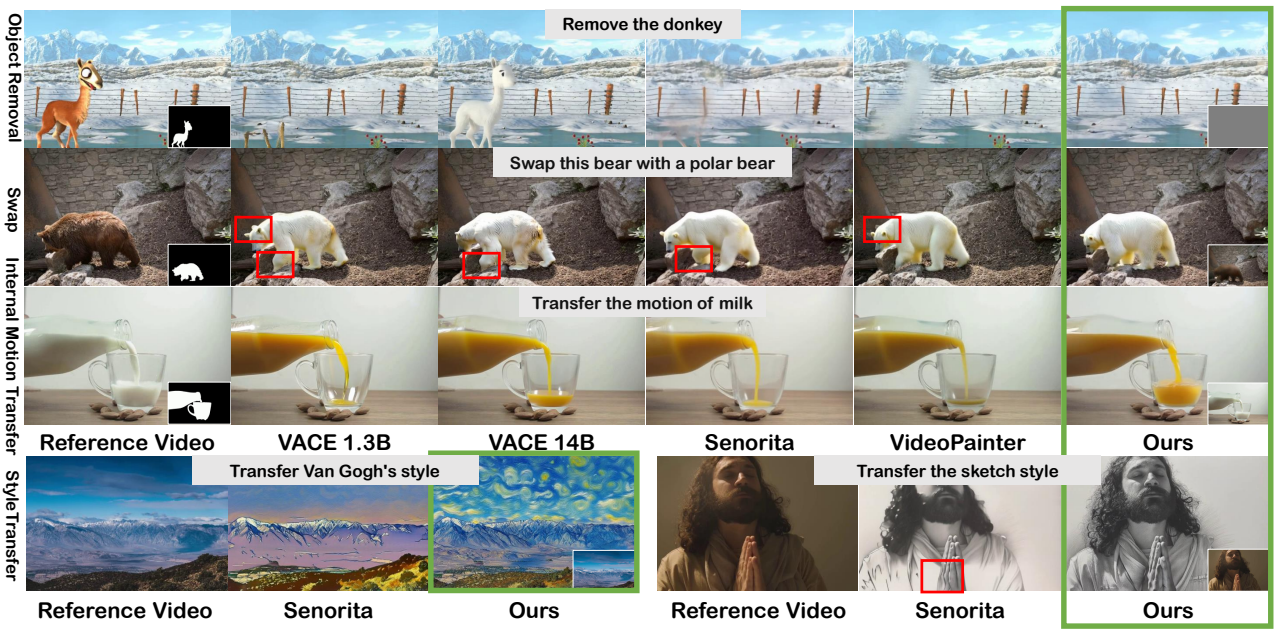

Figure 4: Our O-DisCo-Edit method is compared against other baselines across various tasks, including object removal, swap, object inside motion transfer, and style transfer. The bottom right of our results highlights the A-O-DisCo.

Figure 5: A comparison on the outpainting task, showing the cropped region from the same location within the videos.

Style Transfer. O-DisCo-Edit attains the highest ArtFID, as shown in Tab. 3 (h). In contrast, Senorita exhibits a very low CFSD, which indicates a tendency for its generated videos to align with the original reference content. As depicted in Fig. 4, such alignment is detrimental to style transfer quality. Therefore, our method received higher user study evaluations, demonstrating its superior visual fidelity.

4.3 Ablation Analysis

As shown in Tab. 4, we ablate on O-DisCo-Edit. (1) Comparing row 1 and row 2, a significant improvement in both $PSNR_P$ and $SSIM_P$ is observed with the inclusion of the CFP module. Thus this results validate CFP module's effectiveness in non-edited regions preservation. (2) When contrasting row 3/4 with row 2, both A-O-DisCo and IDP module individually enhance the generated video quality, the appearance consistency of the edited regions, and the non-edited regions preservation capability. (3) A further comparison between rows 3/4 and row 5 reveals that the combination of the A-O-DisCo and IDP modules leads to an even greater improvement in model performance.

Metric	Video Quality		Alignment	Preservation	
Model	TC↑	FVD↓	CLIP-I_E↑	$PSNR_P$↑	$SSIM_P$↑
w/o ①②③	0.9837	887.6	**92.13**	21.57	0.6534
w/o ②③	0.9834	866.3	91.03	25.32	0.8037
w/o ②	<u>0.9839</u>	<u>776.8</u>	91.69	<u>26.09</u>	<u>0.8087</u>
w/o ③	<u>0.9839</u>	796.7	91.60	25.47	0.8053
O-DisCo-Edit	**0.9840**	**711.8**	<u>91.84</u>	**26.25**	**0.8096**

Table 4: Ablation studies on swap task. ① CFP module, ② A-O-DisCo, ③ IDP module. "w/o ①②③" denotes training with R-O-DisCo and inference with a fixed signal, entirely omitting IDP and CFP modules. "w/o ②③" indicates training without IDP module and inference with a fixed signal. "w/o ②" refers to using a fixed inference signal. "w/o ③" indicates training without the IDP module.

5 Conclusion

In this work, we introduced O-DisCo-Edit, a unified framework designed to address the key challenges in controllable video editing. Our core innovation, O-DisCo, unifies various editing signals into a single, noise-based representation. This not only dramatically simplifies the training process and reduces resource demands but also enables multi-granularity editing during inference. Paired with our CFP module, O-DisCo-Edit can accomplish high-fidelity editing while robustly preserving unedited regions. Comprehensive experiments on eight different tasks show that O-DisCo-Edit achieves new SOTA results, outperforming both specialized and multi-task models. This success offers a new perspective on video editing research, that a single, unified control can be both versatile and precise without sacrificing efficiency.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62576191). Furthermore, We are sincerely grateful to the IIGROUP for their valuable feedback and insightful suggestions. We also extend our sincere gratitude to Mr. Haisu Wu for his insightful comments and meticulous revisions, which significantly enhanced the clarity and presentation of this manuscript.

References

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, J.; Xia, M.; Fu, X.; Wang, X.; Mu, L.; Cao, J.; Liu, Z.; Hu, H.; Bai, X.; Wan, P.; et al. 2025. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*.

Bian, Y.; Zhang, Z.; Ju, X.; Cao, M.; Xie, L.; Shan, Y.; and Xu, Q. 2025. Videopainter: Any-length video inpainting and editing with plug-and-play context control. *arXiv preprint arXiv:2503.05639*.

Cai, Q.; Chen, J.; Chen, Y.; Li, Y.; Long, F.; Pan, Y.; Qiu, Z.; Zhang, Y.; Gao, F.; Xu, P.; et al. 2025. HiDream-I1: A High-Efficient Image Generative Foundation Model with Sparse Diffusion Transformer. *arXiv preprint arXiv:2505.22705*.

Chen, C.; Yang, X.; Shu, J.; Wang, C.; and Li, Y. 2025a. LMP: Leveraging Motion Prior in Zero-Shot Video Generation with Diffusion Transformer. *arXiv preprint arXiv:2505.14167*.

Chen, Q.; Ma, Y.; Wang, H.; Yuan, J.; Zhao, W.; Tian, Q.; Wang, H.; Min, S.; Chen, Q.; and Liu, W. 2024. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*.

Chen, W.; Wu, J.; Xie, P.; Wu, H.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023. Control-a-video: Controllable text-to-video generation with diffusion models. *CoRR*.

Chen, Y.; Wang, J.; Liu, L.; Chu, R.; Zhang, X.; Tian, Q.; and Yang, Y. 2025b. O-DisCo-Edit: Object Distortion Control for Unified Realistic Video Editing. *arXiv preprint arXiv:2509.01596*.

Chung, J.; Hyun, S.; and Heo, J.-P. 2024. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8795–8805.

Gu, Z.; Yan, R.; Lu, J.; Li, P.; Dou, Z.; Si, C.; Dong, Z.; Liu, Q.; Lin, C.; Liu, Z.; et al. 2025. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. *arXiv preprint arXiv:2501.03847*.

HaCohen, Y.; Chiprut, N.; Brazowski, B.; Shalem, D.; Moshe, D.; Richardson, E.; Levin, E.; Shiran, G.; Zabari, N.; Gordon, O.; et al. 2024. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Huang, Y.; Wang, W.; Zhao, S.; Xu, T.; Liu, L.; and Chen, E. 2025. Bind-Your-Avatar: Multi-Talking-Character Video Generation with Dynamic 3D-mask-based Embedding Router. *arXiv preprint arXiv:2506.19833*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Jiang, Z.; Han, Z.; Mao, C.; Zhang, J.; Pan, Y.; and Liu, Y. 2025. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*.

Li, Q.; Xing, Z.; Wang, R.; Zhang, H.; Dai, Q.; and Wu, Z. 2025a. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. *arXiv preprint arXiv:2503.16421*.

Li, R.; Xing, D.; Sun, H.; Ha, Y.; Shen, J.; and Ho, C. 2025b. TokenMotion: Decoupled Motion Control via Token Disentanglement for Human-centric Video Generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1951–1961.

Li, X.; Xue, H.; Ren, P.; and Bo, L. 2025c. Diffueraser: A diffusion model for video inpainting. *arXiv preprint arXiv:2501.10018*.

Liang, S.; Yu, Z.; Zhou, Z.; Hu, T.; Wang, H.; Chen, Y.; Lin, Q.; Zhou, Y.; Li, X.; Lu, Q.; et al. 2025. OmniV2V: Versatile Video Generation and Editing via Dynamic Content Manipulation. *arXiv preprint arXiv:2506.01801*.

Lin, G.; Gao, C.; Huang, J.-B.; Kim, C.; Wang, Y.; Zwicker, M.; and Saraf, A. 2023. Omnimatterf: Robust omnimatte with 3d background modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23471–23480.

Liu, C.; Li, R.; Zhang, K.; Lan, Y.; and Liu, D. 2024a. StableV2V: Stablizing Shape Consistency in Video-to-Video Editing. *arXiv preprint arXiv:2411.11045*.

Liu, L.; Liu, Q.; Qian, S.; Zhou, Y.; Zhou, W.; Li, H.; Xie, L.; and Tian, Q. 2024b. Text-animator: Controllable visual text video generation. *arXiv preprint arXiv:2406.17777*.

Pan, T.; Liu, L.; Liu, J.; Zhang, X.; Tang, J.; Wu, G.; and Tian, Q. 2025. RASA: Replace Anyone, Say Anything–A Training-Free Framework for Audio-Driven and Universal Portrait Video Editing. *arXiv preprint arXiv:2503.11571*.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.

Tu, Y.; Luo, H.; Chen, X.; Ji, S.; Bai, X.; and Zhao, H. 2025. Videoanydoor: High-fidelity video object insertion with precise motion control. *arXiv preprint arXiv:2501.01427*.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open

and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, G.; Fan, S.; Liu, H.; Song, Q.; Wang, H.; and Xu, J. 2025a. Consistent Video Editing as Flow-Driven Image-to-Video Generation. *arXiv preprint arXiv:2506.07713*.

Wang, H.; Ouyang, H.; Wang, Q.; Wang, W.; Cheng, K. L.; Chen, Q.; Shen, Y.; and Wang, L. 2025b. Levitor: 3d trajectory oriented image-to-video synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12490–12500.

Wright, M.; and Ommer, B. 2022. Artfid: Quantitative evaluation of neural style transfer. In *Proceedings of DAGM German Conference on Pattern Recognition*, 560–576. Springer.

Wu, C.; Huang, L.; Zhang, Q.; Li, B.; Ji, L.; Yang, F.; Sapiro, G.; and Duan, N. 2021. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*.

Wu, X.; Feng, Z.; Yang, S.; Qin, Y.; Chen, H.; and Liu, Y. 2024. Safety risk perception and control of water inrush during tunnel excavation in karst areas: An improved uncertain information fusion method. *Automation in Construction*, 105421.

Yang, S.; Hou, L.; Huang, H.; Ma, C.; Wan, P.; Zhang, D.; Chen, X.; and Liao, J. 2024a. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *Proceedings of ACM SIGGRAPH 2024 Conference Papers*, 1–12.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Yariv, G.; Kirstain, Y.; Zohar, A.; Sheynin, S.; Taigman, Y.; Adi, Y.; Benaim, S.; and Polyak, A. 2025. Through-The-Mask: Mask-based Motion Trajectories for Image-to-Video Generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 18198–18208.

Ye, Z.; He, X.; Liu, Q.; Wang, Q.; Wang, X.; Wan, P.; Zhang, D.; Gai, K.; Chen, Q.; and Luo, W. 2025. UNIC: Unified In-Context Video Editing. *arXiv preprint arXiv:2506.04216*.

Yin, S.; Wu, C.; Liang, J.; Shi, J.; Li, H.; Ming, G.; and Duan, N. 2023. DragNUWA: Fine-grained Control in Video Generation by Integrating Text. *Image, and Trajectory*.

Zhang, S.; Zhuang, J.; Zhang, Z.; Shan, Y.; and Tang, Y. 2025a. FlexiAct: Towards Flexible Action Control in Heterogeneous Scenarios. *arXiv preprint arXiv:2505.03730*.

Zhang, X.; Duan, Z.; Gong, D.; and Liu, L. 2025b. Training-free motion-guided video generation with enhanced temporal consistency using motion consistency loss. *arXiv preprint arXiv:2501.07563*.

Zhang, X.; Zhou, H.; Qin, H.; Lu, X.; Yan, J.; Wang, G.; Chen, Z.; and Liu, Y. 2025c. Enabling versatile controls for video diffusion models. *arXiv preprint arXiv:2503.16983*.

Zhao, Q.; Ma, Z.; and Zhou, P. 2025. DreamInsert: Zero-Shot Image-to-Video Object Insertion from A Single Image. *arXiv preprint arXiv:2503.10342*.

Zhou, S.; Li, C.; Chan, K. C.; and Loy, C. C. 2023. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10477–10486.

Zi, B.; Peng, W.; Qi, X.; Wang, J.; Zhao, S.; Xiao, R.; and Wong, K.-F. 2025a. MiniMax-Remover: Taming Bad Noise Helps Video Object Removal. *arXiv preprint arXiv:2505.24873*.

Zi, B.; Ruan, P.; Chen, M.; Qi, X.; Hao, S.; Zhao, S.; Huang, Y.; Liang, B.; Xiao, R.; and Wong, K.-F. 2025b. Se\~norita-2M: A High-Quality Instruction-based Dataset for General Video Editing by Video Specialists. *arXiv preprint arXiv:2502.06734*.