FreeMask: Rethinking the Importance of Attention Masks for Zero-Shot Video Editing

Lingling Cai[1]*, Kang Zhao[2], Hangjie Yuan[1], Yingya Zhang[2], Shiwei Zhang[2], Kejie Huang[1]†

[1]Zhejiang University
[2]Tongyi Lab

{cailingling22, hj.yuan, huangkejie}@zju.edu.cn, {zhaokang.zk, yingyazhang.zyy, zhangjin.zsw}@alibaba-inc.com

Abstract

Text-to-video diffusion models have made remarkable advancements. Driven by their ability to generate temporally coherent videos, research on zero-shot video editing using these fundamental models has expanded rapidly. To enhance editing quality, structural controls are frequently employed in video editing. Among these techniques, cross-attention mask control stands out for its effectiveness and efficiency. However, when cross-attention masks are naively applied to video editing, they can introduce artifacts such as blurring and flickering. Our experiments uncover a critical factor overlooked in previous video editing research: cross-attention masks are not consistently clear but vary with model structure and denoising timestep. To address this issue, we propose the metric Mask Matching Cost (MMC) that quantifies this variability and propose **FreeMask**, a method for selecting optimal masks tailored to specific video editing tasks. Using MMC-selected masks, we further improve the masked fusion mechanism within comprehensive attention features, e.g., temp, cross, and self-attention modules. Our approach can be seamlessly integrated into existing zero-shot video editing frameworks with better performance, requiring no control assistance or parameter fine-tuning but enabling adaptive decoupling of unedited semantic layouts with mask precision control. Extensive experiments demonstrate that FreeMask achieves superior semantic fidelity, temporal consistency, and editing quality compared to state-of-the-art methods.

Code — https://freemask-edit.github.io

1 Introduction

With the growing interest in video diffusion models (Wang et al. 2023a; Yuan et al. 2024; Ho et al. 2022), research on video editing based on these foundational models is booming, which aims at high-performance video re-creation. Among these editing paradigms, zero-shot video editing has attracted considerable attention due to its efficiency and low computational cost. Recent zero-shot video editing approaches (Bai et al. 2024; Ku et al. 2024) have leveraged open-source pre-trained text-to-video diffusion models (Wang et al. 2023b; Zhang et al. 2023a) to improve temporal consistency, effectively alleviating issues like visual flickering across frames, which often occurred in earlier image-based methods (Qi et al. 2023; Geyer et al. 2023).

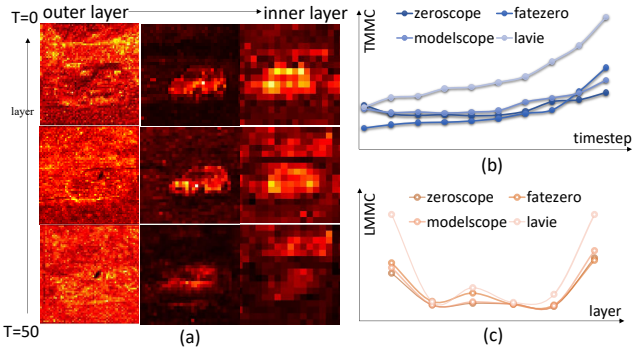

Figure 1: (a) Visualization of the 'jeep' cross-attention maps across layers and denoising timesteps on zeroscope (Cerspense 2023). (b) TMMC of different models across timesteps. (c) LMMC of different models across layers. Refer to the appendix for a detailed description.

To further improve editing quality, auxiliary structural controls are often incorporated into models. These controls help identify specific regions for editing, enabling precise local modifications while preserving the coherence of unedited regions with the original video. However, existing methods often fail to achieve desirable structural controls. Recent related studies can be categorized into two types: **(1)** those relying on external control mechanisms, and **(2)** those relying on internal control mechanisms. For the first type, methods resort to optical flow estimation (Yang et al. 2023; Liang et al. 2024), depth estimation, edge estimation (Ouyang et al. 2024), or semantic segmentation (Kahatapitiya et al. 2024; Esser et al. 2023). These methods have the following drawbacks: **(1)** The above time-agnostic masks may not suit all editing tasks, such as shape editing, where the edited shape does not conform to the source shape mask. **(2)** These methods require external pre-trained models, thereby introducing extra computational overhead for inference. **(3)** Directly applying models designed for image processing to video processing often leads to significant artifacts (Ouyang et al. 2024; Yang et al. 2023), necessitating specialized designs and re-training (Liang et al. 2024). These shortcomings make it a non-optimal choice for

*Work done during the internships at Tongyi Lab
†Corresponding author

video editing. The second type of video editing that relies on internal masks offers a more efficient solution. The iconic method Prompt2prompt (Hertz et al. 2022) has demonstrated the structure control capability of the cross-attention map. Nevertheless, *naively utilizing the imprecise masks derived from cross-attention maps leads to inferior video editing results*, which has been overlooked by previous research.

In response to this issue, we rethink the importance of attention masks in zero-shot video editing and propose **FreeMask** that leverages the attention masks by strategic mask usage. FreeMask is based on *two key observations* illustrated in Fig. 1: **(1)** Cross-attention maps become clearer as denoising progresses. **(2)** Cross-attention maps exhibit the most precision in the middle layer, with the outer layer being too noisy and the inner layer being too imprecise due to low resolution. To quantify these observations, we introduce Mask Matching Cost (MMC), an MIoU-based metric that measures layer-wise and timestep-wise attention mask precision, referred to as LMMC and TMMC, respectively.

Using these two metrics, we further design task-adaptive MMC for mask usage with varying precision control, tailored for different editing tasks, *e.g.*, style translation, attribute editing, and shape editing.

Moreover, most zero-shot video editing heavily relies on the blending of the source and edited attention features (Ku et al. 2024; Qi et al. 2023; Bai et al. 2024). Nevertheless, determining the optimal blending ratio is challenging. Insufficient blending can lead to structural distortion, while excessive blending results in a completely identical video to the original, a common issue leading to prompt misalignment. To address this issue, we integrate MMC-selected masks with masked feature blending across various attention layer types—temporal, cross, and self-attention, as shown in Tab 1. This comprehensive masked fusion typically eliminates the need for ratio selection and significantly enhances the editing accuracy. It is worth noting that FreeMask can be seamlessly integrated into existing zero-shot video editing frameworks without additional control assistance or parameter fine-tuning. The extensive experiments provide evidence of the numerous advantages FreeMask offers for zero-shot video editing without user-specific controls.

2 Related Works

2.1 Video Editing

Video generation methods have shifted from earlier efforts using deep generative models like GANs (Goodfellow et al. 2020; Saito, Matsumoto, and Saito 2017) to diffusion models (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Song, Meng, and Ermon 2020), which produce higher-quality visual outputs. As diffusion models gain popularity, video editing techniques rapidly evolve to harness their potential, which can be broadly categorized into **training-based** and **zero-shot** methods and should satisfy the criteria of fidelity, alignment, and quality (Xing et al. 2023).

Training-based methods enhance temporal consistency and motion fidelity by introducing temporal layers into pre-trained text-to-image (T2I) diffusion models (Rombach et al. 2022). These methods train on large-scale datasets for generalizability or fine-tune specific video instances for efficiency. Models such as GEN-1 (Esser et al. 2023) and VideoComposer (Wang et al. 2024) exemplify the former, while Tune-A-Video (Wu et al. 2023), ControlVideo (Zhang et al. 2023b), and VideoP2P (Liu et al. 2024) represent the latter. Both approaches involve high training costs and may result in spatial detail distortion.

In contrast, **zero-shot methods** avoid extensive training by directly leveraging pre-trained diffusion models. Some early works based on pre-trained T2I models unavoidably introduce spatiotemporal distortion, such as Tokenflow (Geyer et al. 2023), Rerender-A-Video (Yang et al. 2023), CoDef (Ouyang et al. 2024), and FateZero (Qi et al. 2023). While some recent works significantly improved temporal consistency by utilizing open-source conditional video diffusion models like AnyV2V (Ku et al. 2024) and UniEdit (Bai et al. 2024), semantic fidelity remains challenging. Zero-shot methods often struggle to balance semantic fidelity and prompt alignment due to reliance on feature fusion between source and edited videos. To alleviate the restrictive fusion, mask guidance is widely employed.

2.2 Mask Guidance for Video Editing

Mask guidance is an intuitive approach to facilitate strcutural control in video editing. Explicit segmentation masks enhance editing quality, as demonstrated by early works using generative adversarial networks (Wang et al. 2018, 2019) and advanced by recent diffusion-based methods like Text2Live (Bar-Tal et al. 2022), Pix2Video (Ceylan, Huang, and Mitra 2023) and Object-Centric Diffusion (Kahatapitiya et al. 2024). Despite their effectiveness, explicit masks can be resource-intensive and constrain shape modifications.

Alternatively, **cross-attention masks** offer a more flexible solution. Methods like VideoP2P (Liu et al. 2024) and FateZero (Qi et al. 2023) use cross-attention masks inspired by Prompt2Prompt (Hertz et al. 2022) to improve spatiotemporal coherence while reducing reliance on external models. Although efficient, cross-attention masks may still produce artifacts if alignment variations are not addressed carefully.

2.3 Other Structural Guidance for Video Editing

In addition to semantic masks, other structural information such as **depth/edge maps** can aid structural control. ControlNet (Zhang, Rao, and Agrawala 2023) is widely used as a training-free plug-and-play module or integrated into video diffusion models for retraining. For training-free applications, examples include Rerender-A-Video (Yang et al. 2023), ControlVideo (Zhang et al. 2023b), and MoonShot (Zhang et al. 2024). For retraining, examples are VideoComposer (Wang et al. 2024), Gen-1 (Esser et al. 2023), and Control-A-Video (Chen et al. 2023). While training-free methods facilitate structural editing, they often struggle with temporal consistency. Retraining methods improve temporal coherence but are costly. Additionally, techniques like FLATTEN (Cong et al. 2023), Rerender-A-Video (Yang et al. 2023), CoDef (Ouyang et al. 2024), and FlowVid (Liang et al. 2024) use **optical flow** for mask creation or canonical image construction. Overall, while these

structural guidance methods improve video editing by preserving structure, they face challenges with shape modifications and require a balance between editing flexibility and structural accuracy. Aside from structural guidance, some works like FRAG (Yoon et al. 2024) and Slicedit (Cohen et al. 2024) leverage high-frequency components or spatiotemporal slices to enhance editing quality, but this work focuses primarily on structural guidance.

3 Preliminaries

Text-to-video diffusion models. Given a source video $\mathbf{X_0} \in \mathbb{R}^{F\times 3\times H\times W}$ with F frames of RGB images (height H and width W set to 512), a source prompt P_0, and an editing prompt P_1, we aim to generate an edited video $\mathbf{X_1}$ using text-to-video diffusion models. We encode $\mathbf{X_0}$ into a latent $\mathbf{Z_0} \in \mathbb{R}^{F\times d\times h\times w}$ using an image encoder $\mathcal{E}$, where d is the latent dimension and h, w are 64. The latent $\mathbf{Z_0}$ can be decoded back to $\widehat{\mathbf{X}}_\mathbf{0} \approx \mathbf{X_0}$ using a decoder $\mathcal{D}$.

During inference, a 3D-Unet (Çiçek et al. 2016) ϵ_θ denoises the latent variable $\mathbf{Z}_t$ using the noise schedule $\bar{\alpha}_t$ and a text embedding $\psi(P)$ from a text encoder ψ. For DDIM sampling (Song, Meng, and Ermon 2020), the latent variable $\mathbf{Z}_{t-1}$ is updated at timestep t from $\mathbf{Z}_t$ by: $\mathbf{Z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\mathbf{Z}_t-\sqrt{1-\bar{\alpha}_t}\cdot\epsilon_\theta(\mathbf{Z}_t,\psi(P),t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(\mathbf{Z}_t,\psi(P),t)$, where σ_t denotes the added noise, usually set to 0 for deterministic sampling. Repeating this process for T steps yields the final clean latent.

DDIM inversion. The reverse operation of DDIM sampling (Song, Meng, and Ermon 2020), known as DDIM inversion (Mokady et al. 2023), estimates $\mathbf{Z}_{t+1}$ from $\mathbf{Z}_t$: $\mathbf{Z}_{t+1} = \sqrt{\frac{\bar{\alpha}_{t+1}}{\bar{\alpha}_t}}\mathbf{Z}_t + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}}-1}-\sqrt{\frac{1}{\bar{\alpha}_t}-1}\right)\cdot\epsilon_\theta(\mathbf{Z}_t,\psi(P),t)$. Repeating this process for T steps yields the final noisy latent variable $\mathbf{Z}_T$, which is then used as the initial latent for the denoising process.

Cross-attention masks in diffusion models. To ensure the generated video aligns with the text description, the model leverages cross-attention (Wei et al. 2020) during denoising. Specifically, the deep spatial-temporal features of the noisy video frame $\tilde{\mathbf{Z}}_\mathbf{t}$, are projected into a query matrix $Q_C = \ell_{Q_C}(\tilde{\mathbf{Z}}_\mathbf{t}) \in \mathbb{R}^{F\times S\times d'}$, where S is sequence length of Q_C and d' is the embedding dimension. The text embedding is projected into a key matrix $K_C = \ell_{K_C}(\psi(P)) \in \mathbb{R}^{F\times S'\times d'}$ and a value matrix $V_C = \ell_{V_C}(\psi(P)) \in \mathbb{R}^{F\times S'\times d'}$ using learned linear projections $\ell_{Q_C}, \ell_{K_C}, \ell_{V_C}$, where S' is sequence length of K_C and V_C. The *cross-attention maps* are given by:

$$A_C = \text{Softmax}\left(\frac{Q_C {K_C}^T}{\sqrt{d'}}\right) \tag{1}$$

As a whole, $A_C \in \mathbb{R}^{F\times S\times S'}$ reflects the similarity between Q_C and K_C, which enables correlations between text embedding and semantic layout. The binary masks are obtained by thresholding A_C with a constant τ (Avrahami, Lischinski, and Fried 2022; Qi et al. 2023).

4 Methodology

4.1 Method Overview

As illustrated in Fig. 2, we leverage the well-established inversion-then-denoising pipelines (Avrahami, Lischinski, and Fried 2022; Qi et al. 2023; Ku et al. 2024) and cached attention features belonging to 3D-Unet of all inversion steps for editing. In the pre-processing step, we first calculate the model-dependent LMMC and TMMC metrics using the DAVIS video dataset. During video editing, after performing DDIM inversion and before denoising, we compute the semantic-adaptive MMC metrics to select attention masks, as detailed inSec 4.2. During denoising, we apply our masks to guide the fusion of different types of attention features, as detailed in Sec 4.3.

4.2 Mask Matching Cost (MMC)

Key observations. As mentioned above and shown in Fig. 1 (a), we observe that the semantic clarity of cross-attention maps is not constant but varies with model structure and denoising timesteps. Furthermore, this structural-aware and timestep-aware semantic clarity is more pronounced in video diffusion models than in image diffusion models. To quantify these observations, we define the metric Mask Matching Cost (MMC) to identify and select masks with precision control.

Mask candidates extraction. Before quantitative observations, we first generate mask candidates $\mathbf{M} = \left\{M_t^l\right\}_{t=0,l=0}^{T-1,L-1}$, where L represents the number of cross-attention layers in the diffusion U-Net, T denotes the number of sampling steps, and $M_t^l \in \mathbb{B}^{F\times H\times W}$ represents the mask candidate at the l-th layer and timestep t. These candidates are derived from cross-attention maps obtained during the DDIM inversion steps, as illustrated in Fig. 2. Specifically, for each cross-attention map, we extract the submap corresponding to the object word p_0 in the prompt P_0. These maps are then binarized using a threshold hyperparameter τ, as described in (Qi et al. 2023; Avrahami, Lischinski, and Fried 2022), and reshaped to match the resolution of the input video frames, resulting in the final mask candidates $\mathbf{M}$.

Layer-wise and Timestep-wise Mask Matching Cost. To quantify the structural and temporal diversity of cross-attention maps, we introduce the Layer-wise Mask Matching Cost (LMMC) and Timestep-wise Mask Matching Cost (TMMC). We use a video dataset with dense annotations and per-frame ground truth segmentation (DAVIS dataset (Federico Perazzi et al. 2016) for practice). For a given video $\mathbf{X_0}$ from the dataset, we use the corresponding ground-truth segmentation masks as reference masks, denoted as $M_0 \in \mathbb{B}^{F\times H\times W}$. We then compute the Mean Intersection over Union (MIoU) between $\mathbf{M}$ and M_0 to evaluate the matching performance:

$$d_t^l = \frac{M_0 \cap M_t^l}{|M_t^l| + |M_0| - |M_0 \cap M_t^l|}, \tag{2}$$

where d_t^l denotes the MIoU of M_t^l and M_0. Then we define:

$$\mathcal{D}_{LMMC} = \left\{ D_l = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{d_t^l} \,\middle|\, l = 0, 1, \ldots, L-1 \right\} \tag{3}$$

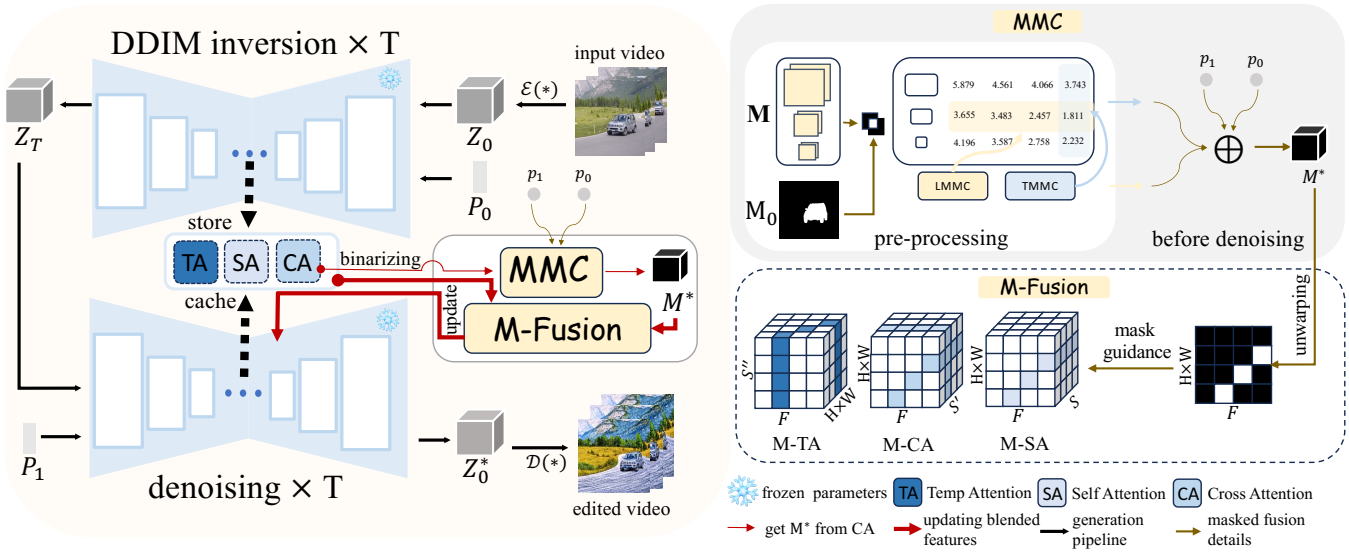

Figure 2: FreeMask overview. FreeMask takes source video $\mathbf{X_0}$ and text prompt P_0 as input. During preprocessing, it stores cross-attention maps for each timestep across all videos in the DAVIS testing dataset to calculate LMMC and TMMC. In the inference stage, $\mathbf{X_0}$ and P_0 are input to DDIM inversion, storing attention features at each timestep and collecting the final latent output as the initial latent for denoising. Before denoising, masks $\mathbf{M}^*$ are adaptively promoted. During denoising, attention features are blended using masks $\mathbf{M}^*$. The final latent output $\mathbf{Z_0^*}$ is then decoded to produce the edited video.

$$l^* = \text{argmin}_{D_l \in \mathcal{D}_{LMMC}} D_l \tag{4}$$

where $\mathcal{D}_{LMMC}$ denotes the set of LMMC across layers, and l^* represents the layer index with minimum LMMC.

Similar to the LMMC, we define the TMMC as:

$$\mathcal{D}_{TMMC} = \left\{ D_t = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{d_t^l} \,\middle|\, t = 0, 1, \ldots, T-1 \right\} \tag{5}$$

$$t^* = \text{argmin}_{D_t \in \mathcal{D}_{TMMC}} D_t \tag{6}$$

where $\mathcal{D}_{TMMC}$ is the set of TMMC across timsteps, and t^* is timestep with minimum TMMC.

Rethinking cross-attention matching for video and text. Using the designed metrics, we quantify our observations and provide a theoretical explanation. We analyzed four T2V models (Lavie (Wang et al. 2023b), modelscope (Wang et al. 2023a), zeroscope (Cerspense 2023), and a pseudo-T2V model of Fatezero (Qi et al. 2023)) with 60 prompt-video pairs and 20 randomly selected video samples with per-frame single instance segmentation masks from DAVIS (Federico Perazzi et al. 2016). We found that the clarity variety of cross-attention masks is common in video diffusion models and systematic at the model level, meaning the regularity of the variety is related to model architecture rather than input videos. The details of the key observations are illustrated in Fig. 1 (b) and (c): (1) *Cross-attention matching accuracy exhibits an inverted U-shape with increasing layers.* This discrepancy stems from that the outer layers contain low-level spatial information while the inner layers capture higher-level spatial information. Text embeddings are high-level information, and the similarity of semantic level is a prerequisite for effective video-text matching, as the cross-attention map calculates the matric product of the query from text embeddings and the key from video latent. Nonetheless, semantic matching accuracy does not equate to mask matching accuracy, as the resolution in the innermost layer is too low, leading to higher LMMC. (2) *Matching accuracy increases with denoising timesteps.* Timesteps in the denoising process run in reverse order, with T decreasing towards 0. Initially, high latent noise results in less accurate low-level spatial information and, consequently, less precise high-level semantic matching.

Task-adaptive Mask Matching Cost. After quantifying the structure-aware and timestep-aware semantic disparity with LMMC and TMMC separately, we select the optimal masks to guide the editing process. Let $\mathbf{M}^* = \left\{ \widehat{M}_t^l \right\}_{t=0,i=0}^{T-1,L-1}$ be the optimal masks, where $\widehat{M}_t^l \in \mathbb{B}^{F \times H \times W}$ represents the optimal mask at the l-th layer and timestep t. We highlight this systematic semantic disparity in cross-attention as an advantage of mask precision control to meet the needs of different tasks. Time-agnostic accurate masks are essential for tasks demanding high structural coherence. Conversely, time-aware (coarse-to-fine) masks are needed for tasks like shape editing that involve structural transformation. While previous work relies on accurate segmentation masks and coarse bounding boxes for mask precision control (Xie et al. 2023), the intrinsic properties of cross-attention can achieve this more efficiently and effectively. To this end, we leverage TMMC and LMMC to design the semantic-adaptive MMC, which enables adaptively choosing time-agnostic and time-aware masks through a designed Kronecker delta function δ :

$$\delta = \begin{cases} 1 & \text{if } p_0 = p_1 \\ 0 & \text{if } p_0 \neq p_1 \end{cases} \tag{7}$$

$$\widehat{M}_t^l = \begin{cases} M_t^{l^*}, & \text{if } \delta = 0, \\ M_{t^*}^{l^*}, & \text{if } \delta = 1. \end{cases} \tag{8}$$

where p_1 corresponds to the object in P_1. For tasks like shape editing, $p_1 \neq p_0$ and time-aware masks are used for flexible shape changes. Otherwise, time-agnostic masks are employed for tasks demanding structural coherence.

4.3 Applications of Mask Guidance

Unlike previous studies that use partially masked fusion for a subset of attention types, as shown in Tab. 1, we apply MMC-selected masks to all major attention types to alleviate blending over-constraint while enhancing editing quality.

Temp-attention feature blending. Temp-attention is crucial for maintaining temporal motion consistency (Bai et al. 2024; Ku et al. 2024). It tracks pixel-level motion by capturing relative transformations between pixels in different frames. However, an excessive fusion of this fine-grained motion information leads to an edited video identical to the source video. To address the over-constraint, we use masks to decouple the edited region within the temp-attention features, where the batch size matches the number of input feature tokens. First, at timestep t of the denoising process, we convert the masks $\mathbf{M_t^*} \in \mathbb{B}^{F \times H \times W}$ into:

$$\mathbf{m_t^*} = \mathcal{F}((1-\delta)\mathbf{M_t^*}) \tag{9}$$

where the mask $\mathbf{M_t^*}$ is reshaped on spatial dimension and flattened into a two-dimensional tensor $\mathbf{m}^* \in \mathbb{B}^{h'w' \times F}$ using the unwrapping function $\mathcal{F}$, where $h' \times w'$ equals to the batch size of temp-attention, which is also the resolution of the image features. For tasks preserving the entire structure like stylization, $\delta = 1$ and $\mathbf{m_t^*} = 0$. We apply $\mathbf{m_t^*}$ to the bath size dimension of temp-attention:

$$\{K_T^*, Q_T^*\} = (1 - \mathbf{m}^*) \odot \{K_T^0, Q_T^0\} + \mathbf{m}^* \odot \{K_T^1, Q_T^1\} \tag{10}$$

where $K_T^0, Q_T^0, K_T^1, Q_T^1, K_T^*, Q_T^*$ represent the source temp-key and temp-query, the edited temp-key and temp-query, and the blended temp-key and temp-query, respectively, all of which belong to the space $\mathbb{R}^{h'w' \times F \times \text{dim}}$, and $\odot$ denotes the Hadamard product.

Cross-attention feature blending. To preserve structural details, cross-attention maps are often reweighted, refined, or replaced with source maps (Hertz et al. 2022; Liu et al. 2024; Qi et al. 2023). Despite the efficacy, the edited cross-attentions can still retain some spatial information about the source object and reduce editing accuracy, e.g., in the prompt "a jeep driving in the countryside," the cross-attention for "a" or "driving" may still show the "jeep" contour, impacting accuracy. To mitigate this issue, we applied additional masked blending for all words in a prompt to enhance standard cross-attention blending operations. We unwarp $\mathbf{M_t^*}$ into $\mathbf{m_t^*}$ with Eq. 9 and blend it with the source cross-attention before performing usual fusion operations:

$$\mathbf{A_C^*} = (1 - \mathbf{m}^*) \odot \mathbf{A_C^0} + \mathbf{m}^* \odot \mathbf{A_C^1} \tag{11}$$

Methods	SM	TM	M-SA	M-CA	M-TA	M-L	Zero-shot	Backbone
Tokenflow	✗	✗	✗	✗	✗	✗	✓	SD
Pix2Video	✗	✗	✗	✗	✗	✗	✓	SD-depth
Rerender-A-Video	✗	✗	✗	✗	✗	✓	✓	ControlNet
Text2Video-Zero	✗	✗	✗	✗	✗	✓	✓	SD
FateZero	✗	✓	✓	✗	✗	✗	w/o shape	SD
CoDeF	✗	✗	✗	✗	✗	✗	✗	ControlNet
Tune-A-Video	✗	✗	✗	✗	✗	✗	✗	SD
Video-P2P	✗	✓	✗	✗	✗	✓	✗	SD
Ours	✓	✓	✓	✓	✓	✓	✓	Any T2V

Table 1: Comparision with different video editing methods. M, SA, CA, TA, and L abbreviate masks, self-attention, cross-attention, temporal attention, and latent, respectively. SM and TM refer to masks that account for structural and timestep-wise variations, respectively.

where $\mathbf{A_C^*}, \mathbf{A_C^0}, \mathbf{A_C^1}$ are the blended, source and edited cross-attentions respectively, belonging to $\mathbb{R}^{F \times h'w' \times S'}$ where S' is equal to the text prompt embedding length.

Self-attention feature blending. We follow the masked self-attention feature blending approach in (Hertz et al. 2022; Qi et al. 2023; Ku et al. 2024), but extend masked blending from original attribute and shape editing to stylization. For stylization, masks are applied inversely, with $\mathbf{m}^*$ masking source self-attention maps. We observe that stylization may lead to undesirable deformation in moving objects, which can be alleviated by masked self-attention blending.

5 Experiments

5.1 Experimental Setup

Datasets and evaluation metrics. We evaluated our method on public DAVIS (Federico Perazzi et al. 2016) videos and 20 Internet videos from pexels (Pexels 2024) depicting various moving subjects. To assess performance, we tested 8-frame videos with a resolution of 512 × 512, using various prompts to produce different editing results for each video.

We report four metrics for quantitative comparison and three metrics for user study. For quantitative comparison, we compute CLIP (Radford et al. 2021) scores for temporal consistency ('Temp') and video-text alignment ('CLIP'), following (Qi et al. 2023). Additionally, we calculate Masked PSNR ('M.PSNR') and LPIPS (Zhang et al. 2018) for structure preservation, following (Liu et al. 2024). For more explanations of metrics, refer to the appendix.

Implementation details. We evaluated our zero-shot video editing approach using the pre-trained T2V model Zeroscope (Cerspense 2023). For each source video, we perform 50 steps of DDIM inversion (Song, Meng, and Ermon 2020), followed by generating outputs using DDIM deterministic sampling (Song, Meng, and Ermon 2020) with classifier-free guidance (Ho and Salimans 2022) at a scale of 7.5. This process takes 1 minute on an NVIDIA A100 GPU. During sampling, attention features are fused with selected masks. The fusion strategies and masks are tailored to specific tasks like stylization, attribute editing and shape editing.

5.2 Comparison of Video Editing

Qualitative comparison results. We compare FreeMask with state-of-the-art methods, including FateZero (Qi et al.

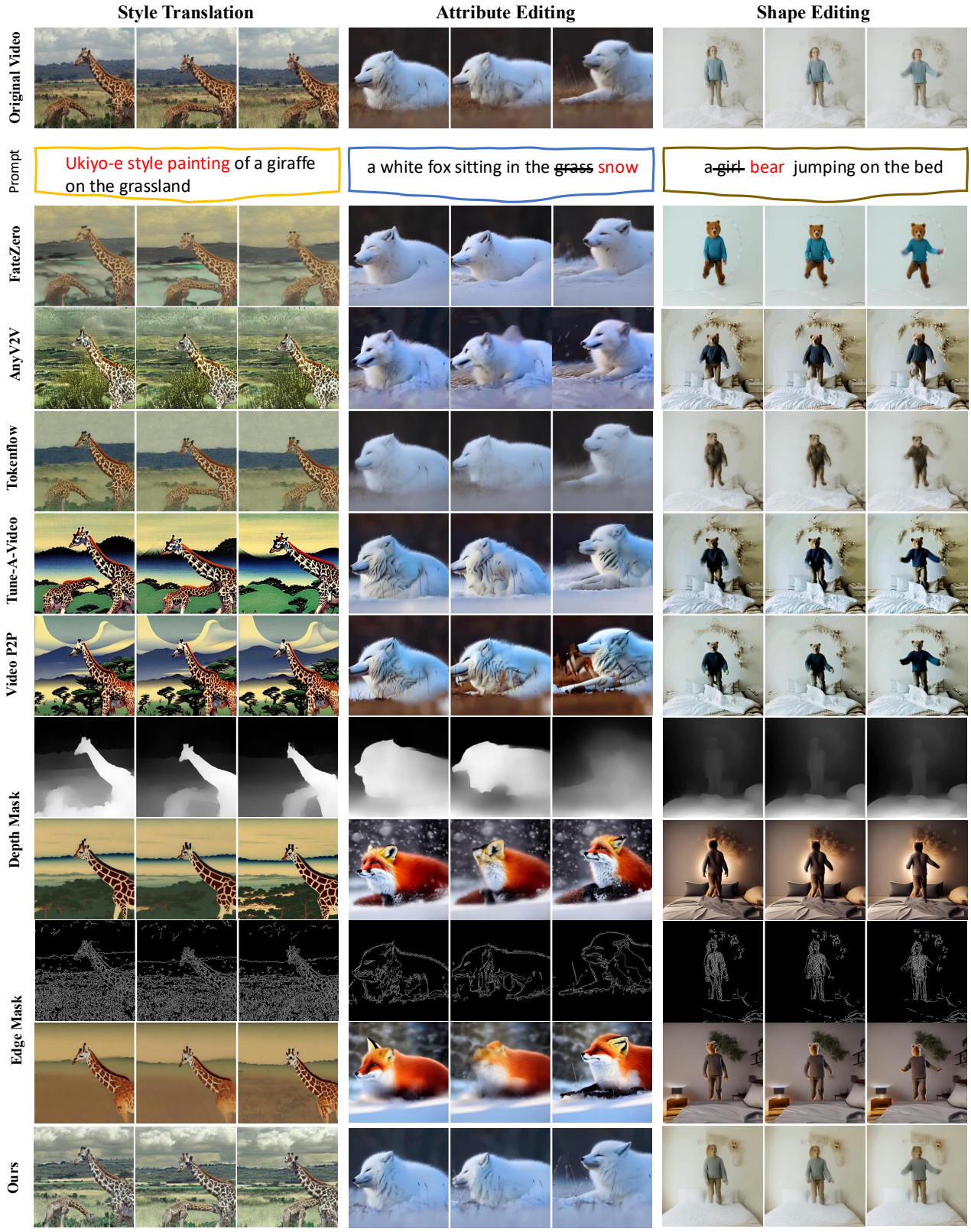

Figure 3: Comparison results with several state-of-the-art approaches on stylization, attribute editing, and shape editing.

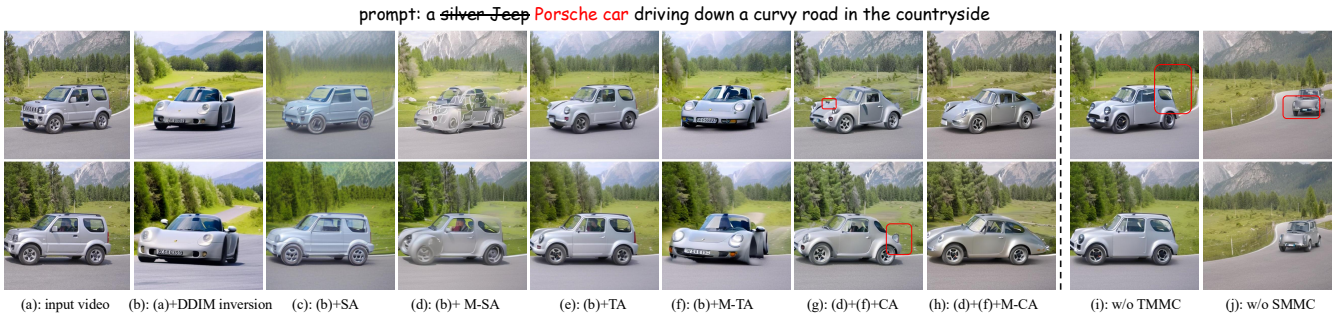

Figure 4: Ablation experiments. The experiments use Zeroscope (Cerspense 2023) as the base model and are conducted on a shape-editing task that changes a Jeep into a Porsche car. In these experiments. (a) is the original input video; (b) shows results using the latent output from DDIM inversion as the initial latent; (c) represents the fusion of self-attention based on (b); (d) involves the fusion of masked self-attention based on (b). Similar logic applies to other sub-captions.

Method	Quantitative Comparison				User Study		
	Temp↑	CLIP↑	M.PSNR↑	LPIPS↓	Edit↑	Image↑	Quality↑
FateZero	0.937	0.278	22.72	0.383	6.12	8.96	7.12
Tokenflow	0.941	0.273	25.41	0.381	6.88	9.01	7.56
Video-P2P	0.923	0.284	18.71	0.508	5.22	5.92	7.35
Tune-A-Video	0.928	**0.285**	17.31	0.514	5.14	5.88	7.43
ControlVideo	0.916	0.281	18.55	0.539	7.99	6.67	7.12
AnyV2V	0.921	0.271	18.23	0.497	5.96	6.74	7.08
FateZero + MMC	0.939	0.279	23.25	0.381	6.55	8.97	7.18
Ours	**0.952**	0.282	**25.94**	**0.366**	**7.68**	**9.23**	**8.90**

Table 2: Quantitative comparison and user study results.

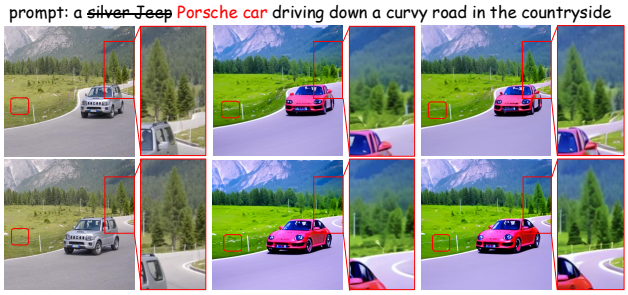

Figure 5: Extension results on shape editing.

2023), TokenFlow (Geyer et al. 2023), ControlVideo (Zhang et al. 2023b) using optimized depth and edge masks, and AnyV2V (Ku et al. 2024) leveraging I2V model. Additionally, we compared it with fine-tuned methods like Tune-A-Video (Wu et al. 2023) and Video-P2P (Liu et al. 2024).

Figure 3 showcases our method's performance across three tasks. For **stylization**, it aligns texture and color with prompts like "Ukiyo-e style painting" while ensuring structural and temporal coherence, unlike others that struggled with content retention. In **attribute editing**, it transforms grass into snowy grass ("snow") while preserving fine details. For **shape editing**, it changes shapes ("bear") while maintaining structure, motion, and expressions, outperforming zero-shot methods on motion consistency and fine-tuned methods on detail preservation.

Quantitative comparison and user study results. As shown in Tab 2, our method significantly outperforms others in quantitative comparison and user studies. Please refer to the appendix for details.

5.3 Ablation Study

Effect of individual cost of MMC. As shown in Fig. 4, (i) demonstrates mask-guided fusion without TMMC, leading to rigid transitions from 'Jeep' to 'Porsche' and unnatural changes, particularly in the tailstock. (j) shows mask-guided fusion without LMMC, resulting in noticeable distortion of structural details, especially in small objects.

Effect of MMC on different fusion mechanisms. We further investigate the design of mask guidance in the fusion of different features. As indicated in Fig. 4, blending masked attentions (including temporal, self, and cross attentions) effectively decouples unedited attributes (dynamic information in temporal attention, structure information in self-attention, and boundary information in cross-attention).

5.4 Extending FreeMask to Other Methods

Qualitative Results. We tested our mask selection using MMC metrics on FateZero (Qi et al. 2023), a zero-shot video editing model based on a flattened text-to-image framework with masked self-attention for shape editing. As shown in Fig. 5, MMC-selected masks enhance structural details, especially around moving objects, while the original masks cause flickering due to suboptimal masking.

Quantitative results. As shown in Tab. 2, we compare FateZero (Qi et al. 2023) before and after adding MMC-selected masks quantitatively and can find it helps with temporal consistency and image editing quality. Additional comparisons, ablation, extension studies and full videos are available in the supplementary materials.

6 Conclusion

We identify a key issue in cross-attention mask guidance for video editing: mask clarity varies across model layers and denoising timesteps. To address this, we propose FreeMask, which uses task-adaptive MMC metrics to select optimal masks, enhancing feature blending while reducing semantic disparity and artifacts and achieving SOTA editing quality.

Acknowledgments

We would like to extend our sincere gratitude to the anonymous reviewers for their invaluable feedback. Additionally, we appreciate the Fundamental Vision Intelligence Team of Tongyi Lab for their generous provision of essential computational resources. This work was supported by Alibaba Research Intern Program.

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18208–18218.

Bai, J.; He, T.; Wang, Y.; Guo, J.; Hu, H.; Liu, Z.; and Bian, J. 2024. Uniedit: A unified tuning-free framework for video motion and appearance editing. *arXiv preprint arXiv:2402.13185*.

Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; and Dekel, T. 2022. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, 707–723. Springer.

Cerspense. 2023. ZeroScope v2 - 576w.

Ceylan, D.; Huang, C.-H. P.; and Mitra, N. J. 2023. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23206–23217.

Chen, W.; Ji, Y.; Wu, J.; Wu, H.; Xie, P.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 424–432. Springer.

Cohen, N.; Kulikov, V.; Kleiner, M.; Huberman-Spiegelglas, I.; and Michaeli, T. 2024. Slicedit: Zero-Shot Video Editing With Text-to-Image Diffusion Models Using Spatio-Temporal Slices. *arXiv preprint arXiv:2405.12211*.

Cong, Y.; Xu, M.; Simon, C.; Chen, S.; Ren, J.; Xie, Y.; Perez-Rua, J.-M.; Rosenhahn, B.; Xiang, T.; and He, S. 2023. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*.

Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; and Germanidis, A. 2023. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7346–7356.

Federico Perazzi, J. P.-T.; McWilliams, B.; Gool, L. V.; Gross, M.; and Sorkine-Hornung, A. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Geyer, M.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2023. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Kahatapitiya, K.; Karjauv, A.; Abati, D.; Porikli, F.; Asano, Y. M.; and Habibian, A. 2024. Object-centric diffusion for efficient video editing. *arXiv preprint arXiv:2401.05735*.

Ku, M.; Wei, C.; Ren, W.; Yang, H.; and Chen, W. 2024. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*.

Liang, F.; Wu, B.; Wang, J.; Yu, L.; Li, K.; Zhao, Y.; Misra, I.; Huang, J.-B.; Zhang, P.; Vajda, P.; et al. 2024. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8207–8216.

Liu, S.; Zhang, Y.; Li, W.; Lin, Z.; and Jia, J. 2024. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8599–8608.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.

Nichol, A.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. *Cornell University - arXiv,Cornell University - arXiv*.

Ouyang, H.; Wang, Q.; Xiao, Y.; Bai, Q.; Zhang, J.; Zheng, K.; Zhou, X.; Chen, Q.; and Shen, Y. 2024. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8089–8099.

Pexels. 2024. Pexels Stock Photos. https://www.pexels.com. [Accessed: 2024-08-05].

Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15932–15942.

Radford, A.; Kim, J.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Amanda, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *Cornell University - arXiv,Cornell University - arXiv*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saito, M.; Matsumoto, E.; and Saito, S. 2017. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, 2830–2839.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. *arXiv: Learning,arXiv: Learning*.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023a. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, T.-C.; Liu, M.-Y.; Tao, A.; Liu, G.; Kautz, J.; and Catanzaro, B. 2019. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Liu, G.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*.

Wang, X.; Yuan, H.; Zhang, S.; Chen, D.; Wang, J.; Zhang, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2024. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2023b. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*.

Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; and Wu, F. 2020. Multi-Modality Cross Attention Network for Image and Sentence Matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7623–7633.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22428–22437.

Xing, Z.; Feng, Q.; Chen, H.; Dai, Q.; Hu, H.; Xu, H.; Wu, Z.; and Jiang, Y.-G. 2023. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*.

Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2023. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, 1–11.

Yoon, S.; Koo, G.; Kim, G.; and Yoo, C. D. 2024. FRAG: Frequency Adapting Group for Diffusion Video Editing. *arXiv preprint arXiv:2406.06044*.

Yuan, H.; Zhang, S.; Wang, X.; Wei, Y.; Feng, T.; Pan, Y.; Zhang, Y.; Liu, Z.; Albanie, S.; and Ni, D. 2024. InstructVideo: instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6463–6474.

Zhang, D. J.; Li, D.; Le, H.; Shou, M. Z.; Xiong, C.; and Sahoo, D. 2024. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827*.

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

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; and Zhou, J. 2023a. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023b. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*.