Zero-to-Hero: Empowering Video Appearance Transfer with Zero-Shot Initialization and Holistic Restoration

Tongtong Su[1,2*], **Chengyu Wang**[2*], **Haipeng Liao**[3†], **Jun Huang**[2], **Dongming Lu**[1†]
[1]Zhejiang University
[2]Alibaba Cloud Computing
[3]NingboTech University
sutongtong@zju.edu.cn, chengyu.wcy@alibaba-inc.com, lhp@nit.zju.edu.cn,ldm@zju.edu.cn

Abstract

Appearance editing according to user needs is a pivotal task in video editing. Existing text-guided methods often lead to ambiguities regarding user intentions and restrict fine-grained control over editing specific aspects of objects. To overcome these limitations, this paper introduces a novel approach named *Zero-to-Hero*, which focuses on reference-based video editing by disentangling the editing process into two distinct problems. It achieves this by first editing an anchor frame to satisfy user requirements as a reference image and then consistently propagating its appearance across the other frames in the video. To achieve accurate appearance propagation, in the first stage of *Zero-to-Hero*, we leverage correspondences within the original frames to guide the attention mechanism, which is more robust than previously proposed optical flow or temporal modules in memory-friendly video generative models, especially when dealing with objects exhibiting large motions. This offers a solid ZERO-shot initialization that ensures both accuracy and temporal consistency. However, intervention in the attention mechanism results in compounded imaging degradation with unknown blurring and color-missing issues. Following the Zero-Stage, our Hero-Stage Holistically learns a conditional generative model for vidEo RestOration. To accurately evaluate appearance consistency, we construct a set of videos with multiple appearances using Blender, enabling a fine-grained and deterministic evaluation. Our method outperforms the best-performing baseline with a PSNR improvement of 2.6 dB.

Code — https://github.com/Tonniia/Zero2Hero

Introduction

Video editing aims to modify the target video according to user demands. One of the most important sub-tasks is appearance editing (Yang et al. 2023, 2024; Wang et al. 2024), in which the structure of the target video frames is preserved while altering the color, texture of the object, or overall style. Previous text-guided video editing methods addressed this task by leveraging pre-trained Text-to-Image (T2I) models, which rely on textual input (i.e., prompts) as the editing guidance signal (Yang et al. 2023, 2024; Feng et al. 2024; Geyer et al. 2023; Wang et al. 2024; Cong et al. 2023; Kara et al. 2024; Liu et al. 2024b). However, ambiguities in text regarding user intentions may limit fine-grained control over the editing results. Therefore, a more practical solution for users to effectively convey their intentions is to explicitly provide a reference image, leading to the *reference-based video editing* task (Ku et al. 2024; Liu et al. 2024a; Ouyang et al. 2024). This task disentangles video editing into two problems: (1) editing a single image as a reference and (2) consistently propagating it to subsequent frames.

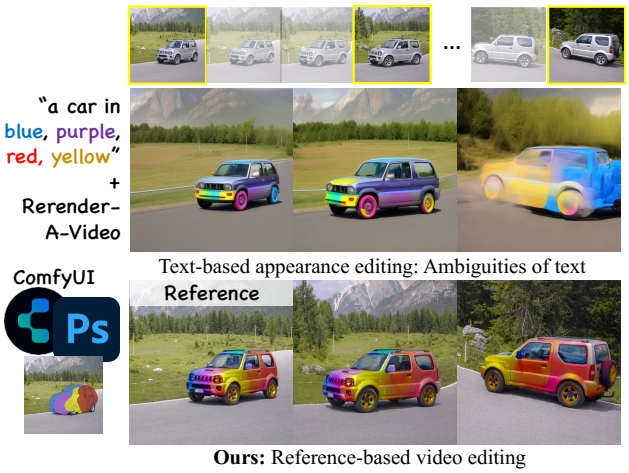

Figure 1: Our reference-based editing method enables users to precisely edit appearances by incorporating complex layouts of color with arbitrary tools such as Photoshop or ComfyUI to create references, then consistently propagate these edits to subsequent frames.

The first sub-task can be addressed using T2I models or arbitrary user manipulation through art design software, allowing for fine-grained appearance changes. The main difficulty lies in the second sub-task: *how to consistently propagate the edited reference frame to other frames*. Current propagation methods can be divided into two groups. The first group uses optical flow obtained from the target video to guide the propagation of reference image features (Yang et al. 2024, 2023; Cong et al. 2023). The performance of these methods can be limited by optical flow estimation (Xu

*Equal contribution.
†Co-corresponding authors.

et al. 2022), which is trained on a specific set of videos. Consequently, its accuracy noticeably degrades when dealing with videos involving significant motion. The second group (Ku et al. 2024; Liu et al. 2024a; Ouyang et al. 2024) leverages Image-to-Video (I2V) models (Zhang et al. 2023; Blattmann et al. 2023) to invert the target video into noisy latent representations, then uses the reference image as a guidance to denoise. However, the video length is constrained by the memory demands of inversion, and the temporal modeling limitations of these memory-friendly I2V models also restrict the range of motion. Recent work (Ouyang et al. 2024) fine-tunes the I2V model with specific target videos. However, for videos with significant motion that deviate far from the I2V domain, it remains challenging to strike a balance between adequately fitting the motion pattern and preventing overfitting of the appearance.

In this work, we explore propagation-based methods but redefine the propagation problem as the more general appearance transfer task (Tumanyan et al. 2022; Park et al. 2020; Mou et al. 2023): maintaining the structure of the target image while utilizing the appearance of the reference image. This will expand the scope of processable videos, eliminating requirements on the range of motion. The task involves finding the semantic correspondence (Ofri-Amar et al. 2023) between reference and target images and then propagating the reference image features into the target ones. Recent approaches connect this task with the self-attention (SA) mechanism in diffusion models (Mou et al. 2024a, 2023; Epstein et al. 2023), leveraging their generative capabilities to support zero-shot appearance transfer. Diffusion models can inherently model intra-similarity for correspondence and simultaneously propagate features using SA. Given two images, expanding SA to cross-image attention (CiA) is a common method for fusing features between images (Chung, Hyun, and Heo 2024; Alaluf et al. 2024; Tewel et al. 2024). However, basic CiA can only capture coarse-grained correspondence, as the query of the target image exhibits similarity to many keys in the reference image (Alaluf et al. 2024). The weighted averaging of matched values leads to a loss of fine-grained details and limits the ability to handle fine-grained appearance transfer. Direct applying contrast value to the attention map can introduce inaccurate transfer since the attention map at the early denoising stage, with a high noise level, cannot represent accurate correspondence. Some research (Tang et al. 2023; Luo et al. 2024; Zhang et al. 2024) has found that DIffusion FeaTures (DIFT) at certain timesteps and U-Net layers can best represent correspondence.

Obtaining correspondence is merely the first step. Directly performing pixel-level swapping based on the highest similarity without incorporating a generative process remains highly sensitive to occasional inaccurate matching, often leading to artifacts, most notably noticeable patch splitting (Zhang et al. 2024). Using correspondence to guide CiA during denoising is more robust, given the output domain constraints of the generative model. However, latent-level swapping in the attention output avoids patch splitting compared to pixel-level swapping but still causes blurring; setting correspondence as an attention mask intervenes in the statistical properties of attention maps, leading to compounded degradation characterized by over-saturated colors and blur patterns of unknown origin (Ahn et al. 2024; Liu et al. 2025). There is an upper limit to what can be achieved with zero-shot methods.

In this paper, we propose *Zero-to-Hero*, which builds upon the aforementioned ZERO-shot intermediate result (i.e., Zero-Stage) as a strong initialization and incorporates the Hero-Stage for Holistic vidEo RestOration. This approach can be formulated as a conditional generation problem (Zhang, Rao, and Agrawala 2023), where the training data pairs require ground truth. The only ground truth available is the reference on the anchor frame. We observed that the Zero-Stage exhibits a consistent pattern of degradation across all frames, indicating that training on the anchor frame has the potential to generalize effectively to all subsequent frames. To accelerate training convergence, the original frame is utilized as an auxiliary condition to encourage shortcut mapping for non-edited regions. To evaluate appearance consistency more accurately beyond semantic-level CLIP-based scores (Huang et al. 2024), we collected a set of 3D objects with multiple appearances and rendered them under significant camera motion to construct videos using *Blender*. This supports fine-grained and deterministic evaluation. Our method outperforms the best-performing baseline with a PSNR improvement of 2.6 dB.

Related Work

Reference-based Video Appearance Editing. Text-guided video editing addresses this task by leveraging pre-trained Text-to-Image (T2I) models, which rely on textual input (i.e., prompts) as the editing guidance signal (Yang et al. 2023, 2024; Geyer et al. 2023; Wang et al. 2024). However, ambiguities in text regarding user intentions may limit fine-grained control over the editing process. Therefore, a more practical solution for users to effectively express their intentions is to use a single image, leading to *reference-based video editing* (Ku et al. 2024; Liu et al. 2024a; Ouyang et al. 2024), which offers a more flexible approach for video editing. AnyV2V (Ku et al. 2024) adopts the earlier memory-friendly I2V model, I2VGen (Zhang et al. 2023), for DDIM inversion (Song, Meng, and Ermon 2020). During the denoising steps, intermediate features are selectively injected to preserve the original motion. This process requires fine-grained hyperparameter tuning of different injection rates at both spatial and temporal modules. Instead of relying on the original temporal module, I2VEdit (Liu et al. 2024a) fine-tunes it to fit the specific target video.

Spatially-Aligned Conditional Generative Models. Spatially-aligned conditional generation (Zhang, Rao, and Agrawala 2023; Mou et al. 2024b; Qin et al. 2023) has been extensively studied in the context of diffusion models. Data collection involves gathering large amounts of images and applying corresponding image processing techniques (e.g., Canny edge detection or depth estimation) to generate conditions, forming training pairs. This can be regarded as obtaining shared-pattern degradation. Various common types of known degradation, such as Gaussian blur (kohya

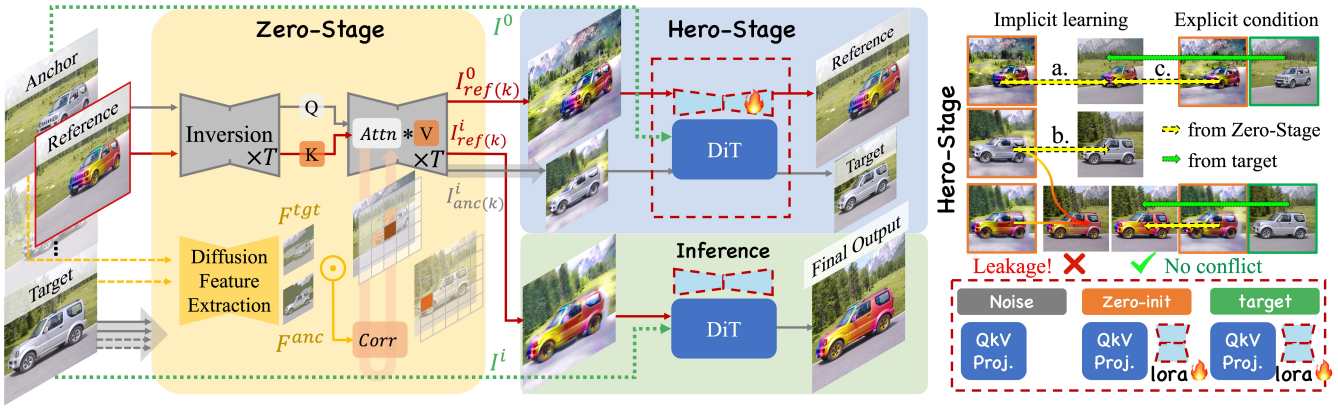

Figure 2: Our framework. **Zero-Stage:** Correspondences ($Corr$) estimated between the anchor and target frames are utilized to guide Cross-image Attention ($Attn$) between the reference and anchor frames, enabling accurate appearance transfer in a zero-shot manner. **Hero-Stage:** We learn a conditional generative model by incorporating LoRA to process conditional tokens. Image pairs serving as potential training data are labeled from (a) to (c) (see Table 1). Target frames as an explicit condition will not conflict with the Zero-initialized condition, whereas implicit learning may lead to the leakage of the target appearance.

ss 2023), tile artifacts (lllyasviel 2023b), and grayscale conversion (lllyasviel 2023a), have corresponding ControlNets. In our problem setting, correspondence as CiA guidance can be regarded as a compounded degradation. Recently, DiT-based diffusion models (Chen et al. 2023; Peebles and Xie 2023; black-forest labs 2023) have demonstrated significant advantages over U-Net models in terms of image quality and prompt understanding. Many works (Tan et al. 2024; Zhang et al. 2025; Tan et al. 2025) fine-tune these base models for conditional generation, demonstrating faster convergence and reduced training data requirements compared to ControlNet.

Method

Given a set of consistent sequences of images from a specific video, we select one anchor frame I^{anc} and edit it exclusively at the appearance level (ensuring spatial alignment, e.g., using ComfyUI with Canny ControlNet) to obtain the reference frame I^{ref}. For each target frame I^{tgt} of the output video, we compose a triplet: $(I^{\text{anc}}, I^{\text{ref}}, I^{\text{tgt}})$. Our goal is to generate the output image frame I^{out}, which depicts the structure present in I^{tgt} while incorporating the appearance edited in I^{ref}. The frame I^{anc} serves as a connection since I^{anc} and I^{ref} are spatially aligned, and the matching between I^{anc} and I^{tgt} is termed correspondence (Zhang et al. 2024; Luo et al. 2024; Tang et al. 2023). In our work, we utilize a pre-trained Stable Diffusion model (Rombach et al. 2022), with VAE encoding the image I into the latent representation z_0, and DDIM inversion (Song, Meng, and Ermon 2020) to obtain the noisy latent z_t. During inversion, attention features in the intermediate steps are preserved. Similar to previous works (Alaluf et al. 2024; Chung, Hyun, and Heo 2024), our method produces an image from a denoising process starting from z_t^{tgt}, with feature injections from the reference image. This process is referred to as Cross-image Attention, which is an extension of Self-Attention. We first review these two mechanisms.

Preliminaries

Self-Attention (SA) and Cross-image Attention (CiA). Self-Attention (SA) serves as a fundamental component in diffusion models for establishing the global structure. The latent z_t is linearly projected into the query (Q), key (K), and value (V) matrices. Attention map $Attn$ is defined as: $Attn = \text{Softmax}\left(\frac{Q \cdot K^{\text{T}}}{\sqrt{d}}\right)$, which computes the similarity among the tokens. The output is defined as the aggregated feature of V weighted by similarity, denoted as $\phi(z_t) = Attn \cdot V$. $Attn$ can represent the structure of a target image when applying DDIM inversion (Song, Meng, and Ermon 2020), while V contains appearance information. During inversion, intermediate $Q^{\text{tgt}}, K^{\text{tgt}}, V^{\text{tgt}}$ are saved and selected for injection into the denoising process for different tasks, e.g., editing (Tumanyan et al. 2023; Hertz et al. 2022) and style transfer (Chung, Hyun, and Heo 2024; Xu et al. 2024).

Cross-image Attention (CiA) extends the concept of Self-Attention (SA) to multiple images. When Q is derived from the target image, and K and V come from a reference image, CiA measures the similarity between tokens from the target (tgt) and reference (ref) images: $Attn = \text{Softmax}\left(\frac{Q^{\text{tgt}} \cdot K^{\text{ref}\text{T}}}{\sqrt{d}}\right)$. This similarity weights the reference V^{ref} to transfer information to the target output: $\phi(z_t) = Attn \cdot V^{\text{ref}}$, which is used in style transfer tasks (Chung, Hyun, and Heo 2024). K^{ref} and V^{ref} can be extended to multiple images, which is beneficial in video processing tasks (Qi et al. 2023; Yang et al. 2023). Although CiA represents similarity and is useful for style transfer, it does not ensure accurate correspondence between I^{ref} and I^{tgt}. The distribution of CiA is scattered, implying that a token in I^{tgt} may interact with numerous tokens in I^{ref}. This interaction averages their V^{ref}, including irrelevant ones, and ultimately leads to the color leakage problem (as shown in the last column of Figure 3). Some works introduce a temperature τ to enhance the contrast of attention maps, encouraging focus on a few patches (Chung, Hyun, and Heo 2024). Oth-

ers boost contrast by increasing the variance of the attention maps (Alaluf et al. 2024). However, CiA still struggles to establish correspondence for spatially unaligned samples in videos with significant motion, making accurate appearance transfer a research challenge.

Correspondence from Diffusion Features. Diffusion models exhibit strong semantic feature extraction capabilities (Zhang et al. 2024; Luo et al. 2024; Tang et al. 2023). These studies investigate which intermediate DIffusion FeaTures (DIFT) are the most effective for establishing semantic correspondence. They add noise at a specific timestep t and feed the noisy latent into the U-Net. Intermediate features from the decoder are extracted through a single denoising step. In our work, we denote intermediate features as F. Similarly to $Attn$ in CiA, the semantic correspondence is based on dot product similarity: $Corr = F^{\text{tgt}} \cdot F^{\text{anc}\mathsf{T}}$. The correspondence between F^{tgt} and F^{anc} is more accurate than that between F^{tgt} and F^{ref}, as both are derived from the original video. Since F^{anc} and F^{ref} are spatially aligned, $Corr$ can be used to guide CiA between F^{tgt} and F^{ref}.

Zero-Stage: Correspondence as CiA Guidance

We propose using correspondence ($Corr$) as accurate guidance for the generative process. There are two potential approaches. The first is latent-level swapping. Compared with pixel-level swapping, we use the intermediate attention module output $\phi(z_t)$ of the reference image and rearrange it according to the top-1 matching index (Ind) in $Corr$:

$$\text{Latent-Swap: } \phi(z_t^{\text{tgt}}) = \text{Swap}(\phi(z_t^{\text{ref}}), \text{Ind}(Corr, 1)). \quad (1)$$

As shown in Figure 3, latent-level swapping eliminates the patch-splitting problem present in direct pixel-level swapping. However, the issue of blurring still persists. The second approach is masked attention. The mask is created by selecting the top-k entries in $Corr$ and setting these positions to 1 in the mask matrix $M(Corr, k)$, where k ranges from 1 to the total number of tokens $h \times w$. These selected positions are assigned large values in the original attention score matrix A, resulting in $A \oplus M(Corr, k)$, before the softmax operation. Using this guided attention and V^{ref} from the reference, a denoising step is:

$$\text{Mask-Attn: } \hat{z}_{t-1}^{\text{tgt}} = \epsilon_\theta(z_t^{\text{tgt}}, A \oplus M(Corr, k), V^{\text{ref}}). \quad (2)$$

The selection of k has a significant impact. As shown in Figure 3, we apply different k values on four frames, including the anchor frame 0. When $k = 1$, the accurate semantic correspondence successfully transfers each part of the car to the target image with corresponding colors. However, the modifications applied to the original attention mechanism are substantial, leading to compounded degradation in image quality, including irregular blurring and color oversaturation. The former is caused by the collapsed structure within the attention map, while the latter arises from the increased feature magnitude due to the low entropy of the identity matrix (Liu et al. 2025). As k increases, image quality gradually returns to normal, while the influence of $Corr$ guidance decreases. When $k = h \times w$, it is equivalent to the original CiA, facing the same problems of color leakage and inaccurate appearance transfer.

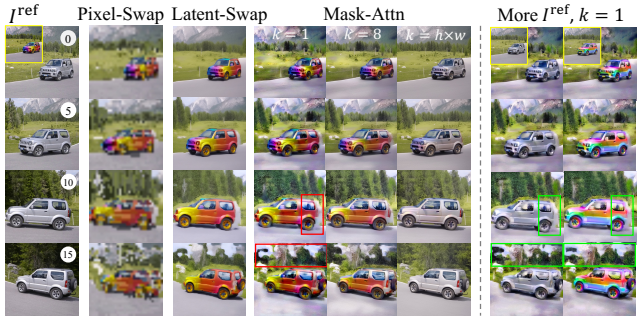

Figure 3: **Left:** Three types of Zero-Stage initialization. **Right:** Using other references results in the same color-missing pattern (red and green boxes).

For both Latent-Swap and Mask-Attn, to achieve precise migration, these zero-shot methods inevitably make some compromises in terms of image quality. We observed that the degradation pattern is similar across different frames. We also have the ground-truth high-quality version of the transferred result on the anchor frame, which is the reference itself, serving as a potential training pair. Therefore, a natural question arises: *can we train a conditional generative model to restore all other frames with compounded degradation?*

Hero-Stage: Zero-Stage as Condition

We use the intermediate result from the Zero-Stage as the image condition and perform Holistic vidEo RestOration, addressing issues such as irregular blurring and color oversaturation. We select one of the Zero-Stage initialization methods arbitrarily; without loss of generality, we choose Mask-Attn with $k = 1$. Given the appearance of the reference, the Zero-Stage output on the anchor frame (frame 0) is denoted as $I^0_{\text{ref}(k)}$. Training on the pair $\{I^0_{\text{ref}(k)}, I^{\text{ref}}\}$ ensures a perfect fit on this anchor frame. We then investigate its ability to generalize to subsequent frames, particularly long-range frames with substantial motion.

Generalization Across Video For target frames that are near the anchor frame 0 (e.g., frame 5 in Figure 3), the degradation patterns are similar to those of the anchor frame. However, for long-range target frames (frames 10 and 15), in addition to the two aforementioned types of degradation, there is also a noticeable color-missing issue caused by unsuccessful matching in significantly changed backgrounds (highlighted in the red box). This color-missing issue does not occur in the training anchor frame pair, so the trained model tends to preserve these missing parts instead of restoring them. Extracting masks for unchanged regions and straightforward replacement of the original target can be sensitive to mask extraction errors. These methods may also fail due to the lack of clear segmentation between objects and the background. To address this, we aim to incorporate the target frames into the training pipeline, automatically providing auxiliary information.

The usage of target frames can be either implicitly learned as labels or explicitly injected as conditions. **Im-**

Implicit	a. zero-ref	$\{I^0_{\text{ref}(k)}, I^{\text{ref}}\}$
	b. zero-tgt	$\{I^i_{\text{anc}(k)}, I^i\}$
Emplicit	c. (zero, tgt)-ref	$\{(I^0_{\text{ref}(k)}, I^0), I^{\text{ref}}\}$

Table 1: Two kind of usage of target frames: implicit learning or explicit condition.

plicit Learning: we encourage the model to implicitly learn unseen regions through the provision of auxiliary training pairs. We use I^{anc} as the reference to construct a series of frames $\{I^0_{\text{anc}(k)}, \cdots, I^n_{\text{anc}(k)}\}$. As shown in Figure 3 (right), frames 10 and 15, with other references, display similar missing regions (in the green box) as when I^{ref} is used as the reference. Specifically, when using I^{anc} as the reference, the corresponding ground truth for each target frame is the frame itself. These auxiliary training pairs are termed zero-tgt: $\{I^i_{\text{anc}(k)}, I^i\}, i = [0, n]$. **Explicit Condition:** we can also explicitly use target frames by introducing an additional conditional branch, denoted as $\{(I^0_{\text{ref}(k)}, I^0), I^{\text{ref}}\}$. Ideally, $I^0_{\text{ref}(k)}$ provides an appearance-transferred intermediate result, while I^0 provides a shortcut to quickly reconstruct the unchanged regions. This training pair is termed (zero, tgt)-ref. The two kinds of usage are summarized in Table 1. We explore which one performs better.

Explicit Condition Can Avoid Leakage As shown in Figure 4, implicit learning with the additional pair b (a+b) achieves better preservation of the car's outline compared to using only pair a, although the missing regions in the background remain unrepaired. With the explicit condition of the target frame (c), the background can be successfully restored. This suggests that using the target frame as an explicit condition can differentiate between edited and non-edited regions: for the edited object regions, it restores the Zero-Stage intermediate results, while for unchanged regions, it directly copies from the target frame as a shortcut.

When the editing involves texture or style changes rather than only color modification (second row, with the reference in watercolor style), implicit learning with training pair b forces the model to learn zero-tgt on each target frame, causing leakage of target appearance. As a result, the style is not successfully changed, and the color layout from the Zero-Stage is not used (see zoom-in on the red box). Conversely, the explicit condition avoids this kind of leakage. This indicates that training the two conditional branches does not cause conflicts, and they can clearly differentiate their respective roles. The appearance of the subsequent frames is successfully transferred to the watercolor style (see zoom-in on the green box).

Few-Shot Conditional Generative Model In our task, we only need to focus on a few-shot images from a single video. The conditional degraded image itself contains sufficient appearance information, making the problem much simpler than training a general conditional model such as those for Canny edges or depth (Zhang, Rao, and Agrawala 2023), which involved high computational cost and slow convergence. Recently, with the development of DiT-based diffusion models, condition injection has become more efficient. These methods (Tan et al. 2024; Zhang et al. 2025) typically convert conditional images into image tokens, concatenate them with noisy latent, and fine-tune DiT modules using LoRA to process conditional tokens. The conditional tokens share the same positional encoding as the corresponding spatially-aligned noisy tokens, which accelerates convergence. Leveraging this efficient fine-tuning architecture along with our solid Zero-Stage initialization, our Hero-Stage achieves faster convergence. Our final method uses the target frame as an explicit condition, making it necessary to employ two independent LoRAs for the model to distinguish between different control purposes.

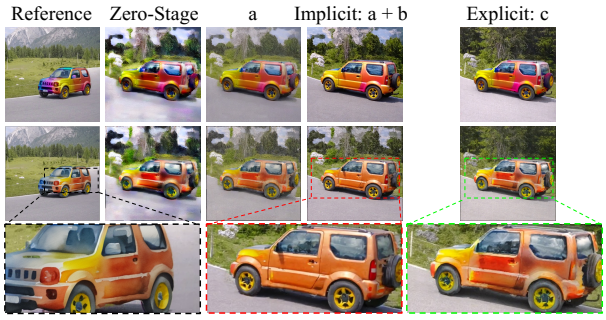

Figure 4: Implicit learning with a + b better preserves the target structure of the car than using only a, but it struggles with style transfer (e.g., watercolor in the second row) and with restoring severely missing background regions. Explicit conditioning with c can solve both problems.

Experiments

Experimental Settings

Datasets We utilize datasets from two sources. The first consists of collections of source videos widely used in video editing (downloaded from (Yang et al. 2024)). These videos have all been uniformly sampled to 16 frames. For each video, we experiment on two sub-tasks. The first sub-task is Colorization, where we convert the target video to grayscale, and the reference is the original anchor frame. The second sub-task is general appearance editing (General-Edit), where the references are processed using Stable Diffusion WebUI with multiple ControlNets for spatial alignment. The editing types include color changes (*a car in red, blue, etc.*), texture changes (*a car made of clay*), and style changes (*a car in watercolor style*). The colorization task supports deterministic evaluation metrics including PSNR, LPIPS, and SSIM for each frame. For General-Edit, there is no ground truth available except for the anchor frame. Therefore, we employ auxiliary metrics to evaluate temporal consistency, including Motion Smoothness (MS) and Subject Consistency (SC), as proposed in VBench (Huang et al. 2024). The colorization task does not cover the general appearance editing task, and General-Edit lacks ground truth, making the evaluation indirect. To address this, we construct a dataset using Blender. We collect five 3D objects, each prepared with

	Colorization			Blender-Color-Edit			General-Edit				
Method	**PSNR (↑)**	**LPIPS (↓)**	**SSIM (↑)**	**PSNR (↑)**	**LPIPS (↓)**	**SSIM (↑)**	**MS (↑)**	**SC (↑)**	**PSNR$_\dagger$ (↑)**	**LPIPS$_\dagger$ (↓)**	**SSIM$_\dagger$ (↑)**
AnyV2V	22.7450	0.1456	0.7703	23.3208	0.1316	0.8174	0.9192	0.8325	26.2379	0.0935	0.8392
I2VEdit	23.4085	0.1231	0.8219	24.1103	0.1317	0.8044	0.9329	0.8724	**27.0925**	0.0830	0.8639
CiA (β^*)	23.2932	0.1486	0.7975	23.0619	0.1377	0.7686	0.9409	**0.9158**	25.5291	0.0943	0.8239
Ours	**28.2063**	**0.0491**	**0.9298**	**26.7640**	**0.0565**	**0.8546**	**0.9428**	0.8978	26.7768	**0.0558**	**0.8886**

Table 2: Results on three appearance editing tasks for all reference-guided methods. For Colorization and Blender-Color-Edit, all frames have ground truth for calculating PSNR (dB), LPIPS, and SSIM. For General-Edit, calculations can only be performed on the anchor frame (†) using the reference as ground truth. Motion Smoothness (MS) and Subject Consistency (SC) are utilized to evaluate temporal consistency.

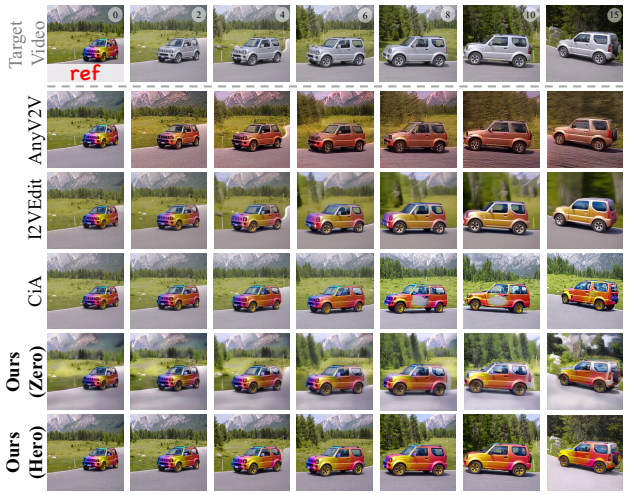

Figure 5: Qualitative results on the General-Edit dataset. Our method maintains the highest consistent fidelity to the reference appearance and target video structure.

three distinct appearances. We employ two classic camera movements: panning and zooming in. The dataset, Blender-Color-Edit, is illustrated in the supplementary material.

Baselines We compare our approach with two reference-based video editing methods, namely AnyV2V (Ku et al. 2024) and I2VEdit (Ouyang et al. 2024), along with an appearance transfer method, CiA with attention contrast (Alaluf et al. 2024) employing Stable Diffusion. AnyV2V directly utilizes I2VGen-XL, applying DDIM inversion and selectively injecting features during the denoising process. It requires varying hyper-parameters for each scenario, including the injection rates at spatial, temporal, and feedforward modules. Following its setting, we adopt an inversion step of 500 and a denoising step of 50, traversing the rates across $[0.2, 0.5, 0.8]$, and selecting the optimal result for each scenario. I2VEdit initially performs LoRA fine-tuning of the temporal module of SVD to adapt to the motion pattern of the target video. We follow its setting by adopting a LoRA rank of $r = 32$. For optimization steps, we employ sufficient steps of $t = 1000$, selecting the best results that balance the fidelity of the motion pattern and reference. The appearance transfer method (Alaluf et al. 2024) employs CiA for semantic matching between two images. It applies a contrast value β to the attention map for more accurate transfer, alongside a guidance value α to diverge from the original appearance. We follow the setting with $\alpha = 3.5$. We observed that the original setting $\beta = 1.67$ may introduce significant structural changes. Thus, we traverse β across $[1, 1.33, 1.67]$ and select the best, referred to as CiA (β^*). Both CiA and our Zero-Stage utilize stable-diffusion-2-1-base, and we use Mask-Attn with $k = 1$ as Zero-Stage for all experiments as the default choice. For our Hero-Stage, we use FLUX.1-dev as base model and perform fine-tuning at a resolution of 512 and an optimization step $t = 400$, utilizing the default LoRA configuration from EasyControl (Zhang et al. 2025), with $r = 128$, $\alpha = 128$.

Comparison with Baselines

Qualitative Results Figure 5 showcases a challenging scenario characterized by substantial object motion, dynamic background changes, and a complex color layout in the reference. AnyV2V exhibits considerable degradation in image quality when tackling such complex motion. I2VEdit adequately fits the motion within 600 steps for this case; however, at this point, the appearance of the reference can no longer effectively guide the subsequent frames. This suggests that for videos with complex motion significantly deviating from the pretrained I2V domain, achieving a balance between motion fitting and reference control in I2V is challenging. The basic CiA method, which applies an optical contrast β^* to attention maps for accurate transfer, alters the structure of the target frame. While this may be acceptable for appearance transfer tasks involving two different objects without strict structure preservation, it is unsuitable for video editing, where maintaining structural consistency of the target frame is crucial. Furthermore, without incorporating more accurate correspondence mechanisms such as DIFT, inaccurate matches can occur, leading to missing color on the car's body. Our Zero-Stage approach with DIFT correspondence guidance, ensures structure preservation and consistent appearance transfer of the car, while the Hero-Stage can further mitigate degradation.

Quantitative Results In Table 2, we quantitatively evaluate the methods across three sub-tasks. For the General-Edit task, our method demonstrates superior performance in Motion Smoothness. CiA, with the optimally searched attention contrast β^*, achieves the highest Subject Consistency but sacrifices preservation of the target structure, resulting in the lowest PSNR on the anchor frame. As illustrated in Figure 5,

	$t = 400/600$			$t = 1000$		
	avg	**anc**	**tgt**$_{(\Delta)}$	**avg**	**anc**	**tgt**$_{(\Delta)}$
tgt-ref	19.38	22.29	$17.52_{(-4.77)}$	22.39	27.09	$18.62_{(-8.47)}$
zero-ref	24.09	29.57	$20.95_{(-8.62)}$	24.59	31.80	$20.92_{(-10.88)}$
Implicit	24.54	28.79	$22.08_{(-6.71)}$	25.77	30.92	$22.91_{(-8.01)}$
Explicit	**26.76**	28.91	$25.51_{(\mathbf{-3.40})}$	**27.38**	29.97	$25.65_{(\mathbf{-4.32})}$

Table 3: Ablation of training pairs on the Blender-Color-Edit dataset: PSNR(↑) of avg: average across all frames; anc: anchor frame; tgt: last target frame; Δ represents the difference between tgt and anc. We evaluate at 400 or 600 steps for one or two conditions, ensuring equal optimization time.

with a changing background, it consistently uses the background from the reference frame rather than preserving the original background. This leads to the highest Subject Consistency score, yet masks issues of appearance inconsistency and structural deformation of the car. Multiple indirect metrics thus need to be evaluated concurrently. Our method attains the best or second-best performance across all metrics. For the Colorization and Blender-Color-Edit tasks, which allow for strict evaluation with ground truth, our method achieves the highest scores.

Ablation Studies

Combinations of Training Data Pairs Two types of target frame usage are presented in Table 1. We also investigate the feasibility of directly learning the transfer from an anchor frame to a reference frame and generalizing this to subsequent target frames, referred to as tgt-ref. We conduct ablation studies using our constructed Blender-Color-Edit dataset to enable accurate evaluation using PSNR. For each scenario, we perform training three times and compute the average to mitigate the effects of randomness.

As shown in Table 3, when directly fine-tuning the model to transfer the original anchor frame to the reference (tgt-ref) with sufficient $t = 1000$ optimization steps, although the PSNR on the anchor frame reaches 27.09, the last target frame remains at 18.62, indicating a complete lack of generalization. With Zero-Stage as initialization, zero-ref achieves a higher PSNR across all target frames on average. However, the generalization ability is limited; even with more optimization steps, the gap between the anchor frame and the last target frame remains substantial ($\Delta = -10.88$). We then further introduce an auxiliary target frame. With Implicit Learning, this gap is reduced to 8.01. With Explicit Condition, the gap is further narrowed to 4.32, reaching the highest average PSNR of 27.38. Moreover, Explicit Condition demonstrates notably faster convergence. Since it involves two conditions, consuming 1.5 times the training time per step compared to methods with one condition, we evaluate it at 400 steps while the others at 600 steps. Explicit Condition achieves the highest average PSNR and the smallest gap between the anchor frame and the last target frame.

Zero-Stage Initialization We emphasize the importance of Zero-Stage initialization. A better initialization can make conditional generation easier and should also remain consistent across frames to enable the anchor-frame learning to generalize to others. As shown in Figure 3, for Zero-Stage, even though Latent-Swap has higher quality than Mask-Attn at the anchor frame, the subsequent frames degrade more noticeably. This might hinder the generalization ability.

Figure 6: Qualitative results on General-Edit. Pixel-Swap initialization leads to the loss of some details in subsequent frames (zoomed-in green and yellow boxes). Latent-Swap achieves higher imaging quality than Mask-Attn at $t = 400$.

Setting	$t = 200$	$t = 400$
Pixel-Swap	22.4982	26.1908
Mask-Attn	**23.7836**	26.7640
Latent-Swap	22.8145	**26.9218**

Table 4: Ablation of Zero-Stage initialization on Blender-Color-Edit. Mask-Attn exhibits faster convergence, while Latent-Swap stabilizes at a higher PSNR in Hero-Stage.

As illustrated in Table 4, during the early optimization steps at $t = 200$, both Latent-Swap and Mask-Attn initializations show a faster increase in PSNR compared to Pixel-Swap initialization and stabilize at a slightly higher average PSNR. This confirms that the quality of Zero-Stage initialization cannot be too low. Mask-Attn demonstrates faster convergence, while Latent-Swap finally stabilizes at a higher PSNR score. This suggests that differences in Zero-Stage quality between frames primarily affect the convergence rate, while the final quality still depends on the overall Zero-Stage quality. As shown in Figure 6, Latent-Swap achieves higher final imaging quality than Mask-Attn.

Conclusion

In this paper, we introduce *Zero-to-Hero* for reference-based video appearance editing. By leveraging accurate correspondence to guide Cross-image Attention, our Zero-Stage ensures fine-grained appearance transfer even for videos with large motion. To address the compounded degradation caused by attention intervention, our Hero-Stage learns a conditional generative model based on training pairs on the anchor frame. This approach generalizes effectively across frames and ensures consistent, holistic restoration of the entire video. Experimental results demonstrate the effectiveness of our method, outperforming baselines in terms of high reference fidelity and temporal consistency.

Acknowledgments

This work is supported by Key Scientific Research Base for Digital Conservation of Cave Temples(Zhejiang University), National Cultural Heritage Administration, and Alibaba Research Intern Program.

References

Ahn, D.; Cho, H.; Min, J.; Jang, W.; Kim, J.; Kim, S.; Park, H. H.; Jin, K. H.; and Kim, S. 2024. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, 1–17. Springer.

Alaluf, Y.; Garibi, D.; Patashnik, O.; Averbuch-Elor, H.; and Cohen-Or, D. 2024. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, 1–12.

black-forest labs. 2023. Flux. [Online] https://github.com/black-forest-labs/flux.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

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.

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*.

Epstein, D.; Jabri, A.; Poole, B.; Efros, A.; and Holynski, A. 2023. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36: 16222–16239.

Feng, R.; Weng, W.; Wang, Y.; Yuan, Y.; Bao, J.; Luo, C.; Chen, Z.; and Guo, B. 2024. Ccedit: Creative and controllable video editing via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6712–6722.

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

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*.

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.

Kara, O.; Kurtkaya, B.; Yesiltepe, H.; Rehg, J. M.; and Yanardag, P. 2024. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6507–6516.

kohya ss. 2023. Gaussian Deblur ControlNet. [Online] https://huggingface.co/kohya-ss/controlnet-lllite.

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*.

Liu, B.; Wang, C.; Su, T.; Ten, H.; Huang, J.; Guo, K.; and Jia, K. 2025. Understanding Attention Mechanism in Video Diffusion Models. *arXiv preprint arXiv:2504.12027*.

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

lllyasviel. 2023a. Colorization ControlNet. [Online] https://huggingface.co/lllyasviel/sd_control_collection/blob/main/ioclab_sd15_recolor.safetensors.

lllyasviel. 2023b. Tile ControlNet. [Online] https://huggingface.co/lllyasviel/ControlNet-v1-1/blob/main/control_v11f1e_sd15_tile.pth.

Luo, G.; Dunlap, L.; Park, D. H.; Holynski, A.; and Darrell, T. 2024. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36.

Mou, C.; Wang, X.; Song, J.; Shan, Y.; and Zhang, J. 2023. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*.

Mou, C.; Wang, X.; Song, J.; Shan, Y.; and Zhang, J. 2024a. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8488–8497.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024b. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4296–4304.

Ofri-Amar, D.; Geyer, M.; Kasten, Y.; and Dekel, T. 2023. Neural congealing: Aligning images to a joint semantic atlas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19403–19412.

Ouyang, W.; Dong, Y.; Yang, L.; Si, J.; and Pan, X. 2024. I2VEdit: First-Frame-Guided Video Editing via Image-to-Video Diffusion Models. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

Park, T.; Zhu, J.-Y.; Wang, O.; Lu, J.; Shechtman, E.; Efros, A.; and Zhang, R. 2020. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33: 7198–7211.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

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.

Qin, C.; Zhang, S.; Yu, N.; Feng, Y.; Yang, X.; Zhou, Y.; Wang, H.; Niebles, J. C.; Xiong, C.; Savarese, S.; et al. 2023. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*.

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.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*.

Tan, Z.; Xue, Q.; Yang, X.; Liu, S.; and Wang, X. 2025. OminiControl2: Efficient Conditioning for Diffusion Transformers. *arXiv preprint arXiv:2503.08280*.

Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2023. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36: 1363–1389.

Tewel, Y.; Kaduri, O.; Gal, R.; Kasten, Y.; Wolf, L.; Chechik, G.; and Atzmon, Y. 2024. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4): 1–18.

Tumanyan, N.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2022. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10748–10757.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Wang, J.; Ma, Y.; Guo, J.; Xiao, Y.; Huang, G.; and Li, X. 2024. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *arXiv preprint arXiv:2406.08850*.

Xu, H.; Zhang, J.; Cai, J.; Rezatofighi, H.; and Tao, D. 2022. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8121–8130.

Xu, Y.; Wang, Z.; Xiao, J.; Liu, W.; and Chen, L. 2024. Freetuner: Any subject in any style with training-free diffusion. *arXiv preprint arXiv:2405.14201*.

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.

Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2024. FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8703–8712.

Zhang, J.; Herrmann, C.; Hur, J.; Polania Cabrera, L.; Jampani, V.; Sun, D.; and Yang, M.-H. 2024. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36.

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

Zhang, Y.; Yuan, Y.; Song, Y.; Wang, H.; and Liu, J. 2025. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*.