Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI, the Arts and Creativity

AdaptEdit: An Adaptive Correspondence Guidance Framework for Reference-Based Video Editing

Tongtong Su[1,2], **Chengyu Wang**[2*], **Bingyan Liu**[3,2], **Jun Huang**[2] and **Dongming Lu**[1*]
[1]Zhejiang University
[2]Alibaba Cloud Computing
[3]South China University of Technology
{sutongtong,ldm}@zju.edu.cn, {chengyu.wcy,huangjun.hj}@alibaba-inc.com,
eeliubingyan@mail.scut.edu.cn

Abstract

Video editing is a pivotal process for customizing video content according to user needs. However, existing text-guided methods often lead to ambiguities regarding user intentions and restrict fine-grained control for editing specific aspects in videos. To overcome these limitations, this paper introduces a novel approach named *AdaptEdit*, which focuses on reference-based video editing that disentangles the editing process. It achieves this by first editing a reference image and then adaptively propagating its appearance across other frames to complete the video editing. While previous propagation methods, such as optical flow and the temporal modules of recent video generative models, struggle with object deformations and large motions, we propose an adaptive correspondence strategy that accurately transfers the appearance from the reference frame to the target frames by leveraging inter-frame semantic correspondences in the original video. By implementing a proxy-editing task to optimize hyperparameters for image token-level correspondence, our method effectively balances the need to maintain the target frame's structure while preventing leakage of irrelevant appearance. To more accurately evaluate editing beyond the semantic-level consistency provided by CLIP-style models, we introduce a new dataset, PVA, which supports pixel-level evaluation. Our method outperforms the best-performing baseline with a clear PSNR improvement of 3.6 dB.

1 Introduction

Video editing is a crucial task for modifying video content based on user needs. Previously, text-guided video editing 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; Yang *et al.*, 2024a; Geyer *et al.*, 2023]. However, ambiguities in text regarding user intentions may limit fine-grained control over the editing process. Therefore, a more practical solution to help users effectively express their intentions is to enable arbitrary video editing based on a single frame. This leads to the *reference-based video editing* task [Ku *et al.*, 2024; Liu *et al.*, 2024a; Ouyang *et al.*, 2024], which disentangles video editing into two problems: (1) editing a single image as a reference and (2) performing reference-image-guided video editing. A simple comparison between text-guided and reference-based video editing is shown in Figure 1.

The first sub-task in reference-based video editing, namely image editing, can be addressed using T2I models or arbitrary user manipulation through art design software. This allows for changes in either the overall style or local, fine-grained color and texture editing. The main difficulty lies in the second sub-task: *how to propagate the edited reference frame to other frames in the video*. Current propagation methods can be divided into three groups. The first group of methods use optical flow obtained from the source video to guide the propagation of reference image features [Yang *et al.*, 2024a; Yang *et al.*, 2023]. The performance of these methods can be limited by the optical flow estimation [Xu *et al.*, 2022], which was trained on rigid body motion in specific types of videos. Consequently, its accuracy noticeably degrades when dealing with shape deformations, perspective shifts, or lighting variations in the video. The second group [Ku *et al.*, 2024; Liu *et al.*, 2024a; Ouyang *et al.*, 2024] leverages recent Image-to-Video (I2V) models such as I2VGen-XL [Zhang *et al.*, 2023] and SVD [Blattmann *et al.*, 2023], using the reference image as a guidance signal. However, the video length and range of motion are restricted due to the temporal modeling limitations of these memory-friendly I2V models when applying DDIM Inversion [Song *et al.*, 2020].

The third group transforms the propagation problem into a more general appearance transfer task [Tumanyan *et al.*, 2022; Park *et al.*, 2020; Mou *et al.*, 2023], which aims to maintain the structure of the target image while utilizing the visual characteristics, such as color and texture, from the reference image. This task involves finding the correspondence between reference and target images and then propagating the reference image features into the target. Recent approaches connect this task with the self-attention (SA) mechanism in diffusion models [Mou *et al.*, 2024; Mou *et al.*, 2023;

*Co-corresponding authors.

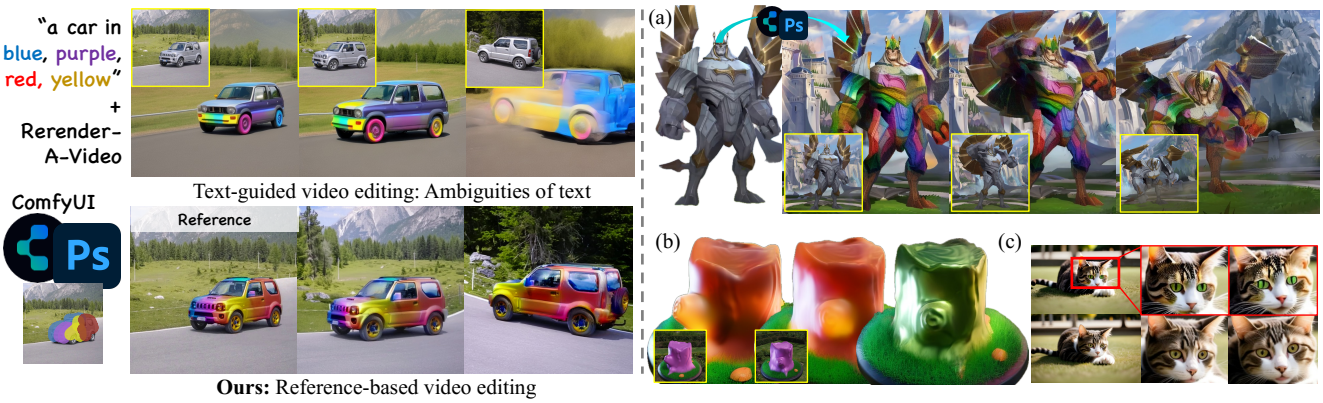

Figure 1: **Left**: Text-guided video editing approaches often fail to adequately capture detailed appearances, as exemplified by Rerender-A-Video. Our reference-based editing method allows users to precisely adjust a single frame according to their needs and then propagate it to the original video (yellow box). **Right**: Typical applications include: (a) re-rendering of large-motion CG, (b) texture modification, and (c) enhancement of generated videos. For comprehensive video results, please see our supplementary material.

Epstein *et al.*, 2023], leveraging their generative capabilities to ensure high image quality and strong generalization for zero-shot performance. 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 *et al.*, 2024; Qi *et al.*, 2023]. 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 effectively. 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 accurately model semantic correspondence. However, these approaches perform pixel-level swapping based on the highest similarity without introducing a generative process, sometimes resulting in artifacts characterized by obvious patch splitting.

In this paper, we propose *AdaptEdit*, a reference-based video editing approach that adaptively leverages accurate DIFT-based correspondence to guide CiA during the generative process of the target video. A basic implementation involves using correspondence as an attention mask, i.e., setting the top-k entries in correspondence to 1 in the mask. The two extreme cases are when $k = 1$ and $k = h \times w$, where $h \times w$ represents the total number of image tokens multiplied by two dimensions. For $k = 1$, the method sometimes overly intervenes in the attention maps and results in artifacts. Conversely, for $k = h \times w$, it corresponds to the original CiA. Naively, we can traverse k to select the most satisfactory results. However, in practice, we find that for different regions in the target image, the optimal k values differ. For regions with high matching confidence, selecting the top-1 is typically sufficient; selecting more can even prove detrimental, as it may lead to unnecessary color leakage from irrelevant parts of the reference image. Therefore, we propose an adaptive correspondence strategy to automatically construct a vector $\mathbf{k}$ with $h \times w$ elements for each region. $\mathbf{k}$ is obtained by formulating a proxy-editing task, where the *paired* pre-editing and ground-truth edited images are generated through deterministic editing, specifically a color-shifting task. Our method surpasses baseline methods in both its fidelity to reference appearances and its temporal consistency. To directly assess the quality of appearance transfer with a focus on fine-grained details, we introduce a new dataset called PVA (Paired Video with Appearance editing). This dataset contains both original and edited versions of target videos rendered from the 3D software Blender, providing ground-truth for pixel-level accuracy evaluation. Our method effectively transfers the reference image's appearance to subsequent frames, surpassing all other appearance transfer baselines in terms of PSNR.

The main contributions of our paper are as follows:

- We propose a reference-based video editing approach named *AdaptEdit*, which allows for the editing of a reference frame according to user intent and subsequently propagates these edits to the remaining target frames.
- In *AdaptEdit*, an adaptive correspondence mechanism is introduced, which leverages precise diffusion feature correspondence to guide the interaction between reference and target frames, demonstrating superior performance over previous appearance transfer baselines.
- We present a novel dataset, PVA (Paired Video with Appearance editing), which comprises both the original videos and their edited counterparts. This pairing provides ground truth support for pixel-level accurate evaluation. Our method surpasses the best-performing baseline, achieving a PSNR improvement of 3.6 dB.

2 Related Works

2.1 Video Editing

Previous works adopt the pre-trained T2I diffusion model for the video editing task [Yang *et al.*, 2023; Geyer *et al.*, 2023; Khachatryan *et al.*, 2023; Qi *et al.*, 2023; Ceylan *et al.*, 2023], utilizing different techniques to ensure consistency between frames. Cross-frame attention [Qi *et al.*, 2023;

Khachatryan *et al.*, 2023; Yang *et al.*, 2023] is limited to ensuring the consistency of global appearance. Rerender-A-Video [Yang *et al.*, 2023] and FRESCO [Yang *et al.*, 2024a] employ optical flow models to warp and fuse latent features. They heavily rely on accurate optical flow estimation, which may not be applicable in regions with rapid motion.

With the rapid development of T2V [Guo *et al.*, 2023; Yang *et al.*, 2024b; Chen *et al.*, 2024a] and I2V [Zhang *et al.*, 2023; Blattmann *et al.*, 2023] models, there has been increasing interest in exploring the integration of these video generative models to process videos as a cohesive whole, thereby ensuring inherent temporal consistency [Ku *et al.*, 2024; Liu *et al.*, 2024a; Shi *et al.*, 2024]. However, the final results are limited by the current model capabilities, which struggle to process videos with large dynamic or complex motion. I2VEdit [Ouyang *et al.*, 2024] fine-tunes I2V model to adapt to specific videos, which requires time-consuming optimization and carries the risk of target video appearance overfitting.

2.2 Appearance Transfer

Appearance transfer aims to transfer the visual characteristics from a reference image to a target image while preserving the structure and layout of the target [Tumanyan *et al.*, 2022; Park *et al.*, 2020; Mou *et al.*, 2023]. Earlier works [Tumanyan *et al.*, 2022; Park *et al.*, 2020; Goel *et al.*, 2023; Chen *et al.*, 2024b] involve a training process, restrict images to specific domains, and require the reference and target images to be spatially aligned. Recent studies have utilized diffusion models in a zero-shot setting without domain restrictions, extending self-attention (SA) to cross-image attention (CiA) [Chung *et al.*, 2024; Khachatryan *et al.*, 2023; Qi *et al.*, 2023]. However, their visual characteristics are limited to overall style, such as color palette and texture features present throughout the image. To achieve a more precise appearance transfer, increasing the contrast of the original scattered elements in CiA can help the target image focus more accurately on the most semantically relevant regions within the reference image [Alaluf *et al.*, 2024; Chung *et al.*, 2024]. For video editing, we have an additional anchor frame: the frame preceding the reference image. Leveraging the semantic correspondence between the anchor and target frames can provide greater accuracy. Our method focuses on effectively utilizing this precise correspondence.

3 *AdaptEdit*: The Proposed Method

Given a target video, we select one anchor frame I^{anc} and edit it using an arbitrary method to obtain the reference frame I^{ref}. We process the video in a frame-wise manner. For each target frame I^{tgt}, we compose a triplet: $(I^{\text{anc}}, I^{\text{ref}}, I^{\text{tgt}})$. Our aim 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} can be viewed as a semantic correspondence task [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 *et al.*, 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 *et al.*, 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.

3.1 Preliminaries and Discussion

Self-Attention (SA) and Cross-Image Attention (CiA)

Self-Attention (SA) serves as a fundamental component in diffusion models for establishing the global structure. Given an input latent z_t comprising $h \times w$ tokens, the intermediate feature in the U-Net $\phi(z_t)$ employs SA linear projections ℓ_q, ℓ_k, and ℓ_v, which map the input image features z_t onto the query, key, and value matrices of a specified dimension d: $Q = \ell_q(\phi(z_t)), \quad K = \ell_k(\phi(z_t)), \quad V = \ell_v(\phi(z_t))$, respectively. The attention map $Attn$ is defined as: $Attn = \text{Softmax}\left(\frac{Q \cdot K^{\mathsf{T}}}{\sqrt{d}}\right)$, which computes the similarity among the tokens. The output is defined as the aggregated feature of V weighted by similarity, denoted as $\phi(z_t) = Attn \cdot V$.

Cross-Image Attention (CiA) extends the concept of 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: $Attn = \text{Softmax}\left(\frac{Q^{\text{tgt}} \cdot K^{\text{ref}\mathsf{T}}}{\sqrt{d}}\right)$. This similarity weights the reference V^{ref} to transfer information to the target output: $Attn \cdot V^{\text{ref}}$. 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}. Some works introduce a temperature τ to enhance the contrast of attention maps, encouraging focus on a few patches [Chung *et al.*, 2024]. Others 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.

Semantic Correspondence in 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}.

One-to-One Matching. Previous works [Zhang *et al.*, 2024] have exploited this correspondence for one-to-one pixel-level

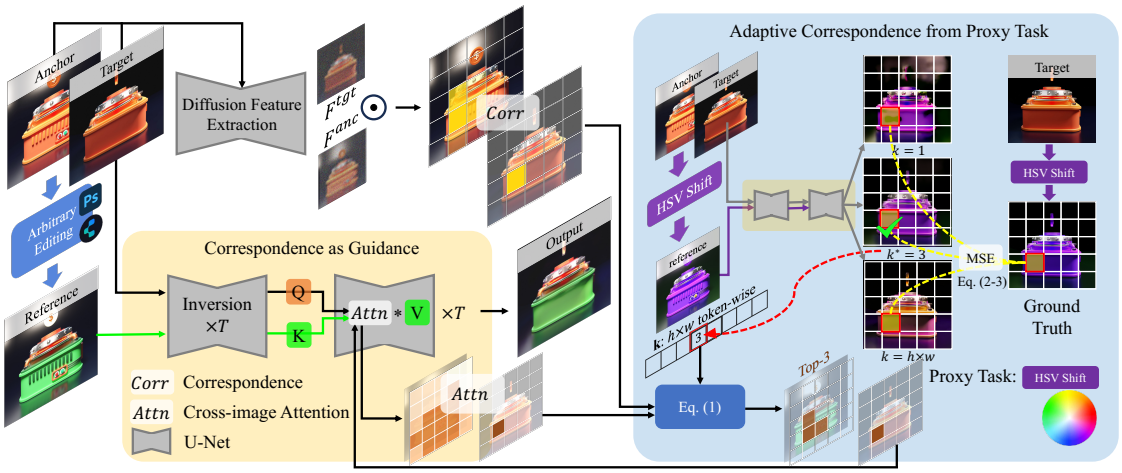

Figure 2: Our *AdaptEdit* framework. Correspondences estimated from the anchor and target frames are sent to the Proxy Task to adaptively assign k for each image token. This adaptive correspondence is then utilized in the CiA mechanism between the reference and anchor frames, guiding the attention to more accurately transfer the appearance from the reference to the target.

swapping, often leading to artifacts with noticeable pixel boundaries due to simplistic operations that fail to integrate correspondence within the diffusion process. Additionally, other studies [Geyer *et al.*, 2023] apply this correspondence as a nearest-neighbor (NN) search in latent space, but still maintain a one-to-one approach. For videos with large motion, such as zooming, a single token is likely to correspond to multiple tokens in the reference.

One-to-All Matching. Although $Corr$ usually outperforms $Attn$ for semantic matching, applying Softmax $\left(\frac{Corr}{\sqrt{d}}\right)$ as a replacement for CiA in the denoising process proves ineffective. This stems from the fact that the attention mechanism cannot be reduced to *Matching and Aggregation* with softmax-similarity weights. Attention standard deviation varies across different denoising timesteps [Chung *et al.*, 2024]. Initially, with noise present, the attention weights are scattered, causing features to interact broadly. Later, as the image clarifies, attention becomes concentrated, allowing each element to engage only with the most similar ones. Consequently, applying $Corr$ from a later denoising stage ($t = 261$) across all timesteps leads to misaligned entropy and degraded imaging quality.

3.2 Adaptive Correspondence as Guidance

We present our adaptive correspondence-based approach for reference-based video editing. As discussed above, the DIFT-based correspondence $Corr$ serves as accurate matching but cannot replace the attention map $Attn$ due to unmatched attention standard deviation. Cross-image attention $Attn$ is compatible with the diffusion denoising process, but it cannot handle fine-grained details due to its inaccuracy.

We propose using $Corr$ as accurate guidance for $Attn$, with the implementation of a masked attention mechanism. 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 places will then be assigned a large value 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 will be:

$$\hat{z}^{\text{tgt}}_{t-1} = \epsilon_\theta(z^{\text{tgt}}_t, A \oplus M(Corr, k), V^{\text{ref}}) \quad (1)$$

where $\hat{z}^{\text{tgt}}_{t-1}$ will be iteratively denoised until obtaining the clean latent $\hat{z}^{\text{tgt}}_0$, thereby achieving appearance transfer.

When determining the value of k, we consider two extreme cases. As illustrated in Figure 3, row 1, we change the body color of the object from orange to green. When $k = 1$, the accurate correspondence can successfully transfer reference values to the target for most regions in the image. However, for some challenging regions, the top-1 selection risks unsuccessful matching, which can result in noticeable artifacts (illustrated by the black hole highlighted in the red box). This artifact may be due to incorrect matching between the shadowed area and the background. By gradually increasing k, artifacts are progressively removed, and when $k = 16$, the black hole artifact is successfully eliminated. However, for other regions that were previously successful when $k = 1$, a larger k might introduce unnecessary correspondence, seemingly obtaining values from irrelevant elements (highlighted in the red circle, indicating color leakage from other parts). When $k = h \times w$, the outcome tends to resemble a color blending of two images rather than an accurate appearance transfer for each region.

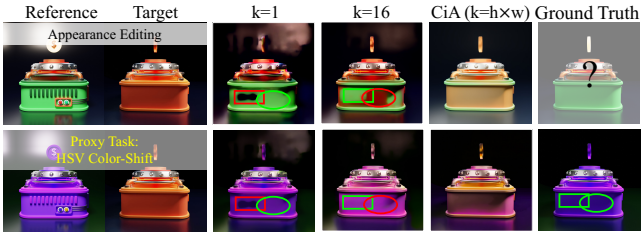

Figure 3: **Row 1**: Increasing k from 1 to $k = h \times w$ reduces artifacts (red box) but may introduce irrelevant color leakage (red circle). **Row 2**: HSV color shift shows consistent pixel-level performance, making it suitable as a proxy for constructing adaptive $\mathbf{k}$.

Therefore, it is crucial to adaptively determine the appropriate k for each region in the target image, i.e., each image token. This involves constructing the vector $\mathbf{k}$ in an adaptive manner. One can traverse a list of k values and visually evaluate which is optimal for each token. Our goal is to automatically construct this through a proxy task.

Adaptive Correspondence from Proxy Task

The proxy task should satisfy three key criteria: **Determinism**: the ground-truth transferred frames are deterministically obtainable; **Alignment**: the proxy task should be highly aligned with the appearance transfer task; and **Efficiency**: the proxy task should be computationally efficient. Next, we describe the proxy task that satisfies these three criteria.

Determinism. We appoint *color-shifting* [Reinhard *et al.*, 2001; Yang *et al.*, 2025] as the proxy task, where the ground truth (GT) after the transfer can be obtained through deterministic image processing algorithms. We should further ensure that the specific type of color-shifting that is most aligned with our final appearance transfer task, i.e., $\mathbf{k}$ determined by the proxy task, is applicable to our final task.

Alignment. The color-shift should show significant differences from the original video frames; otherwise, it may not be sensitive enough to accurately capture the critical points where color leakage begins. We use the HSV transformation on the original frames to construct proxy pairs: $\{z_0^{\text{anc}}, z_0^{\text{tgt}}\} \xrightarrow{\text{HSV}} \{z_0^{\text{ref}}, z_0^{\text{GT}}\}$. Our goal is to identify an optimal vector $\mathbf{k}$ that minimizes the mean square error, ensuring that the denoised z_t^{tgt}, after color is transferred from V^{ref}, closely matches the ground-truth HSV-transferred z_t^{GT}. Specifically, for each image token i, the predicted transferred target is:

$$\hat{z}_{t-1}^{\text{pred}}[i] = \epsilon_\theta(z_t^{\text{tgt}}, A \oplus M(\text{Corr}, \mathbf{k}[i]), V^{\text{ref}})[i]. \quad (2)$$

Here, $\hat{z}_{t-1}^{\text{pred}}[i]$ will be iteratively denoised until a clean latent $\hat{z}_0^{\text{pred}}[i]$ is obtained. The optimization is entry-wise, optimizing $\mathbf{k}$ to minimize the mean square error:

$$\underset{\mathbf{k}[i]\in[1,h\times w]}{\arg\min} \left\| z_0^{\text{GT}}[i] - \hat{z}_0^{\text{pred}}[i] \right\|_2^2. \quad (3)$$

It is worth noting that the HSV transformation uses three parameters: Hue (type of color), Saturation (purity of the color), and Value (brightness of the color), which describe colors as points in a three-dimensional space. A hue shift of 180 degrees is the most discriminative proxy task, which can most accurately capture the critical $\mathbf{k}$ for each image token.

Efficiency. To achieve the most accurate optimization of Eq. 3, the final denoised $\hat{z}_0^{\text{pred}}$ typically requires a setting of $T = 20$ denoising steps. However, this is very time-consuming, as each option for $\mathbf{k}$ necessitates a complete sequence of T denoising steps. In practice, our goal is to quickly assess whether obvious artifacts appear or unnecessary content leakage occurs when k is improper. Therefore, we can tolerate some inaccurate predictions by setting a smaller T. This significantly improves the algorithm's efficiency while resulting in a negligible performance difference.

Another factor affecting efficiency is the search space of $\mathbf{k}$. When the range is set from $[1, h \times w]$, the search space becomes excessively large, especially when processing a video.

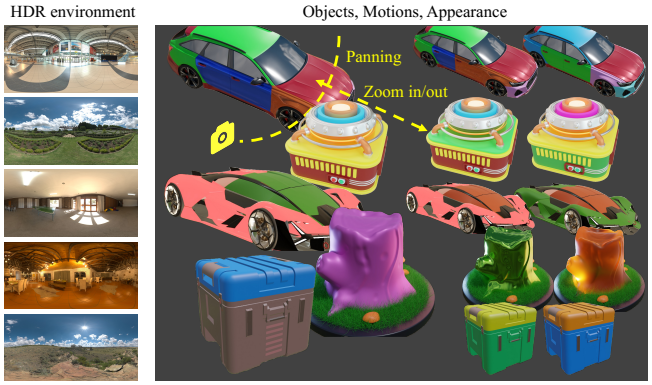

Figure 4: The PVA dataset: 5 objects with 2 appearance variants, spanning 3 motion types and 5 HDR environments.

Between $k = 1$ and $k = h \times w$, the color-shifting error initially decreases as artifacts are removed, then subsequently increases due to color leakage. The minimal point, such as $k = 16$ in Figure 3, where the black hole is just being compensated, is also a critical point denoted as k^*. The search space between $[k^*, h \times w]$ can be skipped, as further increasing would only lead to increased leakage in sensitive regions, and essential information for other regions will be fully captured with $k = h \times w$. Our final search space for $\mathbf{k}$ can thus be narrowed down to $\{1, k^*, h \times w\}$.

3.3 PVA: Proposed Dataset

A main difficulty in video editing evaluation is the lack of ground truth, i.e., pairs of videos before and after editing. Hence, current methods [Huang *et al.*, 2024; Liu *et al.*, 2024b] primarily evaluate the output video based on indirect metrics. Motion smoothness aims to evaluate the temporal consistency between adjacent frames through interpolation. However, the accuracy of the interpolation may diminish with larger motions. Therefore, it should primarily serve as an auxiliary tool for evaluation. Subject consistency utilizes CLIP-based similarity to evaluate temporal consistency; however, its semantic-level evaluation may not be sufficient to capture the fine-grained consistency required for detailed appearance transfer tasks.

If we possess ground-truth pairs of videos before and after editing, we can assess pixel-level accuracy using metrics such as PSNR and LPIPS. Our proposed new dataset is referred to as Paired Video with Appearance editing (PVA). As shown in Figure 4, we collect five 3D objects, each of which allows for the editing of specific component appearances using 3D software such as Blender. For each object, we apply two types of appearance modifications as references. For example, in the first row, we modify the car's appearance. In the first modification, we change the color of the car door from deep blue to orange. In the second modification, we completely swap the colors of each component of the car. We collect five HDR environments as backgrounds, to further assess the robustness of our method under varying lighting conditions. For the types of motion, we utilize three classic camera movements: panning, zooming in, and zooming out.

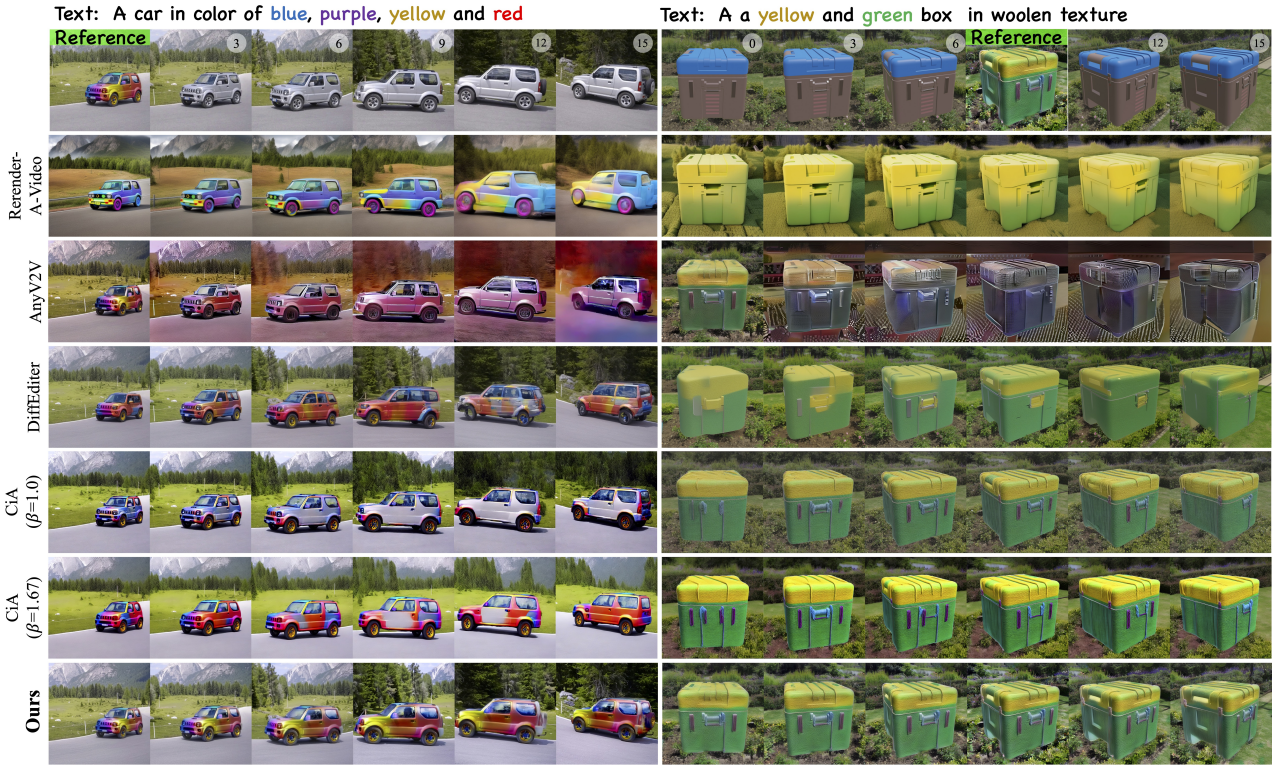

Figure 5: **Qualitative Comparison of *AdaptEdit* with State-of-the-Art Methods.** *AdaptEdit* excels in handling complex appearance descriptions better than text-guided approaches. It surpasses other reference-based baselines by maintaining higher fidelity to reference appearances and target video motion, while offering superior temporal consistency. For more detailed video results, see our supplementary material.

4 Experiments

4.1 Experimental Settings

Datasets. We utilize datasets from two sources. **Multi-Reference-Editing (MRE)** is a collection of videos used in baselines [Yang *et al.*, 2024a], such as *car-turn* (see Figure 1). We edit the anchor frame in each video into various references, ranging from simple overall appearance alterations (e.g., a red car in winter), to complex colorations (e.g., a car in *specified colors*). For creating reference-based edits, we use ComfyUI and Photoshop to modify the anchor frame, producing a satisfactory reference. Additionally, we construct the **MRE-color** dataset by applying deterministic color transformations to the original videos, either converting them to grayscale or applying a 180-degree hue shift, where the reference is the original anchor frame and the ground truth remains the original video. We also leverage **PVA**, our proposed dataset described in Section 3.3.

Baselines. We compare our method with three categories of baselines. The first category is optical-flow based, exemplified by Rerender-A-Video [Yang *et al.*, 2023] and FRESCO [Yang *et al.*, 2024a]. The second category is represented by AnyV2V [Ku *et al.*, 2024], which is designed for reference-based video editing tasks. It edits the first frame and leverages I2VGEN-2 [Zhang *et al.*, 2023] to propagate changes to other frames. The final category encompasses methods focused on appearance transfer tasks. DiffEditor [Mou *et al.*, 2024] requires an additional mask, and both the reference image and target image should be spatially aligned. TokenFlow [Geyer *et al.*, 2023] refers to the method of selecting the top-1 correspondence as the Nearest Neighbor (NN). While the original TokenFlow is text-guided, we adapt its concept for reference-based video editing, naming our adaptation TF-Ref. CiA [Alaluf *et al.*, 2024] refines the original attention map to enhance focus and improve transfer quality. It has two essential hyperparameters: $\alpha = 3.5$, the appearance guidance scale, directs the noisy latent code toward denser regions of the distribution that match the reference appearance while steering it away from the original appearance. β is a contrast operation designed to enhance the variance of the attention maps, encouraging them to focus on more concentrated regions. The optimal β depends on the case, with $\beta = 1.67$ as default and $\beta = 1.0$ indicating no contrast. Our experiments explore these two settings.

Method	MS (↑)	PSNR (↑)	LPIPS (↓)
Target Video	0.9422		
Rerender-A-Video	0.8826	-	-
FRESCO	0.8762	-	-
AnyV2V (inj. 0.5)	0.9324	18.2293	0.3148
AnyV2V (inj. 1.0)	0.9320	19.4697	0.2372
TF-Ref ($k = 1$)	0.9062	17.4829	0.2657
CiA ($k = h \times w$, $\beta = 1.0$)	0.9288	26.2781	0.1239
CiA ($k = h \times w$, $\beta = 1.67$)	0.9154	22.1831	0.1796
AdaptEdit	0.9311	**26.9483**	**0.1213**

Table 1: Results on the MRE dataset. "-" represents text-guided editing that only supports the Motion Smoothness (MS) metric. PSNR and LPIPS are evaluated on MRE-color subset with ground truth.

	PSNR (↑)	LPIPS (↓)	Time (↓)
DiffEditor	19.2673	0.1824	-
CiA ($k = h \times w, \beta = 1.0$)	21.5348	0.1384	-
CiA ($k = h \times w, \beta = 1.67$)	17.0240	0.2350	-
TF-Ref ($k = 1$)	19.2139	0.2339	25.35
$k = 32$	21.9362	0.1353	25.35
$k = h \times w$	21.4857	0.1307	25.35
Ours (HSV, 20, $\{1, k^*, h \times w\}$)	**25.1736**	**0.1065**	+8.97
HSV, 20, $\{1, 2^n, h \times w\}$	24.7081	0.1097	+37.93
HSV, 20, $\{1, h \times w\}$	24.8560	0.1107	+8.03
HSV, 4, $\{1, h \times w\}$	24.2425	0.1219	+2.45
HSV, 1, $\{1, h \times w\}$	20.4482	0.2068	+1.80
Gray, 20, $\{1, h \times w\}$	22.2180	0.1267	+8.03

Table 2: Results on the PVA dataset. Our *AdaptEdit* approach with optimal search space achieve the best PSNR and LPIPS.

4.2 Experimemtal Results

Qualitative Results. Figure 5 highlights the challenges posed by textual ambiguities for the text-guided Rerender-A-Video, which struggles with color consistency. Similarly, AnyV2V, using an I2V model, fails to preserve appearance across frames with complex movements, degrading post-first-frame quality. DiffEditor's requirement for spatial alignment complicates maintaining object shape and appearance in objects with large motion. Both CiA and our approach enhance cross-image attention for appearance transfer. However, only our method consistently and accurately transfers complex appearances across frames. CiA without attention contrast ($\beta = 1.0$) remains sparse attention maps, resulting in uncolored regions on the car (left, frames 9 and 12) and box detail changes (right, all frames). With attention contrast ($\beta = 1.67$), the coloring of the car improves, but inconsistencies and uncolored areas persist. Global contrast adjustments using a constant value will disrupt the original image, leading to structural deformations and color distortions.

Quantitative Results. We evaluate the methods quantitatively on MRE and PVA datasets. Motion Smoothness (MS) [Huang *et al.*, 2024] is determined by interpolating frames [Li *et al.*, 2023] at times $t-1$ and $t+1$, and calculating error with frame t. Table 1 demonstrates that text ambiguities cause methods such as Rerender-A-Video and FRESCO to fail in maintaining temporal consistency, thereby lowering the MS score. Our method surpasses TF-Ref and CiA in MS scores among reference-based techniques. Although AnyV2V reports high MS scores, it contradicts the qualitative results in Figure 5, indicating substantial frame-reference discrepancies. This discrepancy occurs because MS captures only adjacent frame consistency, failing to detect long-term inconsistencies. Therefore, PSNR and LPIPS evaluations on the MRE-color and PVA datasets are necessary for further assessment. As shown in Tables 1 and 2, *AdaptEdit* consistently excels across both datasets, achieving superior PSNR and LPIPS results. On the challenging PVA dataset, there is a notable improvement in PSNR (+3.6 dB) and LPIPS (-0.03) compared to the best-performing baseline, CiA.

Ablation Study. In this section, we emphasize the importance of adaptively constructing $\mathbf{k}$ when applying correspondence to guide CiA. Figure 6 illustrates a case from the PVA dataset. When $k = 1$, the color transfer for each component of the car is correct; however, there is significant blurring in the background and at the edges of details. As k increases, the blurring is resolved, leading to an increase in PSNR (as shown in Table 2, ranging from $k = 1$ to 32). During this process, color transfer gradually becomes less accurate, resulting in an increasing amount of unnecessary content leakage.

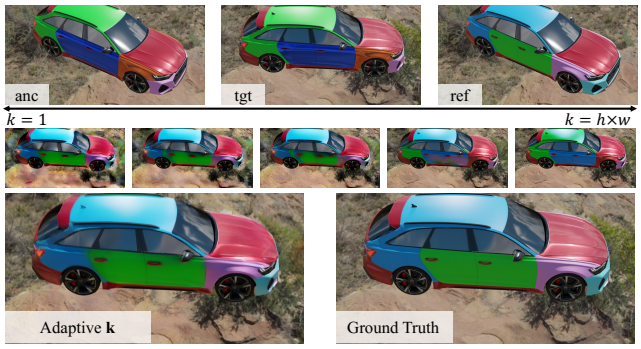

Figure 6: Case study on the importance of the adaptive k. As k increases from 1 to $h \times w$, blurring is reduced in some regions, but content leakage becomes apparent in others. Employing an adaptive k can achieve consistently high quality across the entire image.

We employ our proposed *AdaptEdit* to adaptively construct the vector $\mathbf{k}$ for each image token, according to the proxy task in Eq. 3. Initially, we limit the search to just the two extreme cases, $\{1, h \times w\}$. There is a noticeable improvement in PSNR (+3.2 dB) and LPIPS (-0.03) compared with a fixed k. When we define the search space for $\mathbf{k}$ to span the entire range of $[1, h \times w]$, specifically by traversing all the powers of 2 within this range ($\{1, 2^n, h \times w\}$), the additional time required for the proxy task becomes unacceptable. As shown in Table 2 (where $+$ represents the additional time cost), traversal incurs an extra time cost that is 1.5 times the total cost of all other procedures. Moreover, traversal does not lead to an improvement in PSNR; instead, it can result in degradation. This suggests a slight misalignment between the proxy task and the final task, where an overly fine-grained search space makes the results highly sensitive to small differences in adjacent values of k. By employing an early stop at k^* within a search space of $\{1, k^* = 32, h \times w\}$, we can achieve the best appearance transfer performance efficiently. When decreasing the total denoising timesteps from 20 to 4, there is minimal impact on performance, while significantly enhancing the efficiency of the proxy task. Furthermore, the HSV transformation proved to be the most effective proxy task. Replacing it with grayscale transformation led to a noticeable decrease in PSNR from 24.85 dB to 22.21 dB.

5 Conclusions

In this paper, we introduced *AdaptEdit* for reference-based video editing, which disentangles the complex video editing process by editing a reference frame and propagating its appearance through the video frames. By leveraging an adaptive correspondence strategy, *AdaptEdit* dynamically apply semantic correspondence on cross-image attention, with hyperparameters optimized by a proxy task that ensures determinism, alignment and efficiency. The experimental results demonstrate the effectiveness of *AdaptEdit*, which outperforms the baselines in terms of high reference fidelity and video temporal consistency.

Acknowledgments

This work is supported by Key Scientific Research Base for Digital Conservation of Cave Temples (Zhejiang University), State Administration for Cultural Heritage, and Alibaba Research Intern Program. Work done during T. Su's internship at Alibaba Cloud Computing. Correspondence to: C. Wang and D. Lu.

References

[Alaluf *et al.*, 2024] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

[Blattmann *et al.*, 2023] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[Ceylan *et al.*, 2023] Duygu Ceylan, Chun-Hao P. Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23206–23217, October 2023.

[Chen *et al.*, 2024a] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

[Chen *et al.*, 2024b] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602, 2024.

[Chung *et al.*, 2024] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. 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*, pages 8795–8805, 2024.

[Epstein *et al.*, 2023] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.

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

[Goel *et al.*, 2023] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *CoRR*, 2023.

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

[Huang *et al.*, 2024] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[Khachatryan *et al.*, 2023] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[Ku *et al.*, 2024] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

[Li *et al.*, 2023] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.

[Liu *et al.*, 2024a] Chang Liu, Rui Li, Kaidong Zhang, Yunwei Lan, and Dong Liu. Stablev2v: Stablizing shape consistency in video-to-video editing. *arXiv preprint arXiv:2411.11045*, 2024.

[Liu *et al.*, 2024b] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024.

[Luo *et al.*, 2024] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[Mou *et al.*, 2024] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8488–8497, 2024.

[Ouyang *et al.*, 2024] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[Park *et al.*, 2020] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33:7198–7211, 2020.

[Qi *et al.*, 2023] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023.

[Reinhard *et al.*, 2001] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[Shi *et al.*, 2024] Fengyuan Shi, Jiaxi Gu, Hang Xu, Songcen Xu, Wei Zhang, and Limin Wang. Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7393–7402, 2024.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

[Tumanyan *et al.*, 2022] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022.

[Xu *et al.*, 2022] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022.

[Yang *et al.*, 2023] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.

[Yang *et al.*, 2024a] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8703–8712, 2024.

[Yang *et al.*, 2024b] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[Yang *et al.*, 2025] Yixin Yang, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Colormnet: A memory-based deep spatial-temporal feature propagation network for video colorization. In *European Conference on Computer Vision*, pages 336–352. Springer, 2025.

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

[Zhang *et al.*, 2024] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024.