FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation

Shuai Yang[1*]   Yifan Zhou[2]   Ziwei Liu[2]   Chen Change Loy[2 ✉]

[1]Wangxuan Institute of Computer Technology, Peking University   [2]S-Lab, Nanyang Technological University

williamyang@pku.edu.cn   {yifan006, ziwei.liu, ccloy}@ntu.edu.sg

Figure 1. Our framework enables high-quality and coherent video translation based on pre-trained image diffusion model. Given an input video, our method re-renders it based on a target text prompt, while preserving its semantic content and motion. Our zero-shot framework is compatible with various assistive techniques like ControlNet, SDEdit and LoRA, enabling more flexible and customized translation.

Abstract

*The remarkable efficacy of text-to-image diffusion models has motivated extensive exploration of their potential application in video domains. Zero-shot methods seek to extend image diffusion models to videos without necessitating model training. Recent methods mainly focus on incorporating inter-frame correspondence into attention mechanisms. However, the soft constraint imposed on determining where to attend to valid features can sometimes be insufficient, resulting in temporal inconsistency. In this paper, we introduce **FRESCO**, intra-frame correspondence alongside inter-frame correspondence to establish a more robust spatial-temporal constraint. This enhancement ensures a more consistent transformation of semantically similar content across frames. Beyond mere attention guidance, our approach involves an explicit update of features to achieve high spatial-temporal consistency with the input video, significantly improving the visual coherence of the resulting translated videos. Extensive experiments demonstrate the effectiveness of our proposed framework in producing high-quality, coherent videos, marking a notable improvement over existing zero-shot methods.*

* Work done when Shuai Yang was RAP at S-Lab, NTU.

1. Introduction

In today's digital age, short videos have emerged as a dominant form of entertainment. The editing and artistic rendering of these videos hold considerable practical importance. Recent advancements in diffusion models [33, 34, 36] have revolutionized image editing by enabling users to manipulate images conveniently through natural language prompts. Despite these strides in the image domain, video manipulation continues to pose unique challenges, especially in ensuring natural motion with temporal consistency.

Temporal-coherent motions can be learned by training video models on extensive video datasets [6, 18, 38] or finetuning refactored image models on a single video [25, 37, 44], which is however neither cost-effective nor convenient for ordinary users. Alternatively, zero-shot methods [4, 5, 11, 23, 31, 41, 47] offer an efficient avenue for video manipulation by altering the inference process of image models with extra temporal consistency constraints. Besides efficiency, zero-shot methods possess the advantages of high compatibility with various assistive techniques designed for image models, *e.g.*, ControlNet [49] and LoRA [19], enabling more flexible manipulation.

Existing zero-shot methods predominantly concentrate

on refining attention mechanisms. These techniques often substitute self-attentions with cross-frame attentions [23, 44], aggregating features across multiple frames. However, this approach ensures only a coarse-level global style consistency. To achieve more refined temporal consistency, approaches like Rerender-A-Video [47] and FLATTEN [5] assume that the generated video maintains the same inter-frame correspondence as the original. They incorporate the optical flow from the original video to guide the feature fusion process. While this strategy shows promise, three issues remain unresolved. **1) Inconsistency.** Changes in optical flow during manipulation may result in inconsistent guidance, leading to issues such as parts of the foreground appearing in stationary background areas without proper foreground movement (Figs. 2(a)(f)). **2) Undercoverage.** In areas where occlusion or rapid motion hinders accurate optical flow estimation, the resulting constraints are insufficient, leading to distortions as illustrated in Figs. 2(c)-(e). **3) Inaccuracy.** The sequential frame-by-frame generation is restricted to local optimization, leading to the accumulation of errors over time (missing fingers in Fig. 2(b) due to no reference fingers in previous frames).

To address the above critical issues, we present FRamE Spatial-temporal COrrespondence (**FRESCO**). While previous methods primarily focus on constraining *inter-frame temporal correspondence*, we believe that preserving *intra-frame spatial correspondence* is equally crucial. Our approach ensures that semantically similar content is manipulated cohesively, maintaining its similarity post-translation. This strategy effectively addresses the first two challenges: it prevents the foreground from being erroneously translated into the background, and it enhances the consistency of the optical flow. For regions where optical flow is not available, the spatial correspondence within the original frame can serve as a regulatory mechanism, as illustrated in Fig. 2.

In our approach, FRESCO is introduced to two levels: attention and feature. At the attention level, we introduce FRESCO-guided attention. It builds upon the optical flow guidance from [5] and enriches the attention mechanism by integrating the self-similarity of the input frame. It allows for the effective use of both inter-frame and intra-frame cues from the input video, strategically directing the focus to valid features in a more constrained manner. At the feature level, we present FRESCO-aware feature optimization. This goes beyond merely influencing feature attention; it involves an explicit update of the semantically meaningful features in the U-Net decoder layers. This is achieved through gradient descent to align closely with the high spatial-temporal consistency of the input video. The synergy of these two enhancements leads to a notable uplift in performance, as depicted in Fig. 1. To overcome the final challenge, we employ a multi-frame processing strategy. Frames within a batch are processed collectively,

Figure 2. Real video to CG video translation. Methods [47] relying on optical flow alone suffer (a)(f) inconsistent or (c)(d)(e) missing optical flow guidance and (b) error accumulation. By introducing FRESCO, our method addresses these challenges well.

allowing them to guide each other, while anchor frames are shared across batches to ensure inter-batch consistency. For long video translation, we use a heuristic approach for keyframe selection and employ interpolation for non-keyframe frames. Our main contributions are:

- A novel zero-shot diffusion framework guided by frame spatial-temporal correspondence for coherent and flexible video translation.
- Combine FRESCO-guided feature attention and optimization as a robust intra-and inter-frame constraint with better consistency and coverage than optical flow alone.
- Long video translation by jointly processing batched frames with inter-batch consistency.

2. Related Work

Image diffusion models. Recent years have witnessed the explosive growth of image diffusion models for text-guided image generation and editing. Diffusion models synthesize images through an iterative denoising process [17]. DALLE-2 [33] leverages CLIP [32] to align text and images for text-to-image generation. Imagen [36] cascades diffusion models for high-resolution generation, where class-free guidance [29] is used to improve text conditioning. Stable Diffusion builds upon latent diffusion model [34] to denoise at a compact latent space to further reduce complexity.

Text-to-image models have spawned a series of image

manipulation models [2, 16]. Prompt2Prompt [16] introduces cross-attention control to keep image layout. To edit real images, DDIM inversion [39] and Null-Text Inversion [28] are proposed to embed real images into the noisy latent feature for editing with attention control [3, 30, 40].

Besides text conditioning, various flexible conditions are introduced. SDEdit [27] introduces image guidance during generation. Object appearances and styles can be customized by finetuning text embeddings [8], model weights [14, 19, 24, 35] or encoders [9, 12, 43, 46, 48]. ControlNet [49] introduces a control path to provide structure or layout information for fine-grained generation. Our zero-shot framework does not alter the pre-trained model and, thus is compatible with these conditions for flexible control and customization as shown in Fig. 1.

Zero-shot text-guided video editing. While large video diffusion models trained or fine-tuned on videos have been studied [1, 6, 7, 10, 13, 15, 15, 18, 26, 37, 38, 42, 44, 51], this paper focuses on lightweight and highly compatible zero-shot methods. Zero-shot methods can be divided into inversion-based and inversion-free methods.

Inversion-based methods [22, 31] apply DDIM inversion to the video and record the attention features for attention control during editing. FateZero [31] detects and preserves the unedited region and uses cross-frame attention to enforce global appearance coherence. To explicitly leverage inter-frame correspondence, Pix2Video [4] and TokenFlow [11] match or blend features from the previous edited frames. FLATTEN [5] introduces optical flows to the attention mechanism for fine-grained temporal consistency.

Inversion-free methods mainly use ControlNet for translation. Text2Video-Zero [23] simulates motions by moving noises. ControlVideo [50] extends ControlNet to videos with cross-frame attention and inter-frame smoothing. VideoControlNet [20] and Rerender-A-Video [47] warps and fuses the previous edited frames with optical flow to improve temporal consistency. Compared to inversion-based methods, inversion-free methods allow for more flexible conditioning and higher compatibility with the customized models, enabling users to conveniently control the output appearance. However, without the guidance of DDIM inversion features, the inversion-free framework is prone to flickering. Our framework is also inversion-free, but further incorporates intra-frame correspondence, greatly improving temporal consistency while maintaining high controllability.

3. Methodology

3.1. Preliminary

We follow the inversion-free image translation pipeline of Stable Diffusion based on SDEdit [27] and ControlNet [49], and adapt it to video translation. An input frame I is first mapped to a latent feature $x_0 = \mathcal{E}(I)$ with an Encoder $\mathcal{E}$. Then, SDEdit applies DDPM forward process [17] to add Gaussian noise to x_0

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (1)$$

where $\bar{\alpha}_t$ is a pre-defined hyperparamter at the DDPM step t. Then, in the DDPM backward process [17], the Stable Diffusion U-Net ϵ_θ predicts the noise of the latent feature to iteratively translate $x'_T = x_T$ to x'_0 guided by prompt c:

$$x'_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}\hat{x}'_0 + \frac{(1 - \bar{\alpha}_{t-1})(\sqrt{\alpha_t}x'_t + \beta_t z_t)}{1 - \bar{\alpha}_t}, \quad (2)$$

where α_t and $\beta_t = 1 - \alpha_t$ are pre-defined hyperparamters, z_t is a randomly sampled standard Guassian noise, and $\hat{x}'_0$ is the predicted x'_0 at the denoising step t,

$$\hat{x}'_0 = (x'_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x'_t, t, c, e))/\sqrt{\bar{\alpha}_t}, \quad (3)$$

and $\epsilon_\theta(x_t, t', c, e)$ is the predicted noise of x'_t based on the step t, the text prompt c and the ControlNet condition e. The e can be edges, poses or depth maps extracted from I to provide extra structure or layout information. Finally, the translated frame $I' = \mathcal{D}(x'_0)$ is obtained with a Decoder $\mathcal{D}$. SDEdit allows users to adjust the transformation degree by setting different initial noise level with T, *i.e.*, large T for greater appearance variation between I' and I. For simplicity, we will omit the denoising step t in the following.

3.2. Overall Framework

The proposed zero-shot video translation pipeline is illustrated in Fig. 3. Given a set of video frames $\mathbf{I} = \{I_i\}_{i=1}^N$, we follow Sec. 3.1 to perform DDPM forward and backward processes to obtain its transformed $\mathbf{I'} = \{I'_i\}_{i=1}^N$. Our adaptation focuses on incorporating the spatial and temporal correspondences of $\mathbf{I}$ into the U-Net. More specifically, we define temporal and spatial correspondences of $\mathbf{I}$ as:
- **Temporal correspondence**. This inter-frame correspondence is measured by optical flows between adjacent frames, a pivotal element in keeping temporal consistency. Denoting the optical flow and occlusion mask from I_i to I_j as w_i^j and M_i^j respectively, our objective is to ensure that I'_i and I'_{i+1} share w_i^{i+1} in non-occluded regions.
- **Spatial correspondence**. This intra-frame correspondence is gauged by self-similarity among pixels within a single frame. The aim is for I'_i to share self-similarity as I_i, *i.e.*, semantically similar content is transformed into a similar appearance, and vice versa. This preservation of semantics and spatial layout implicitly contributes to improving temporal consistency during translation.

Our adaptation focuses on the *input feature* and the *attention module* of the decoder layer within the U-Net, since decoder layers are less noisy than encoder layers, and are more semantically meaningful than the x_t latent space:
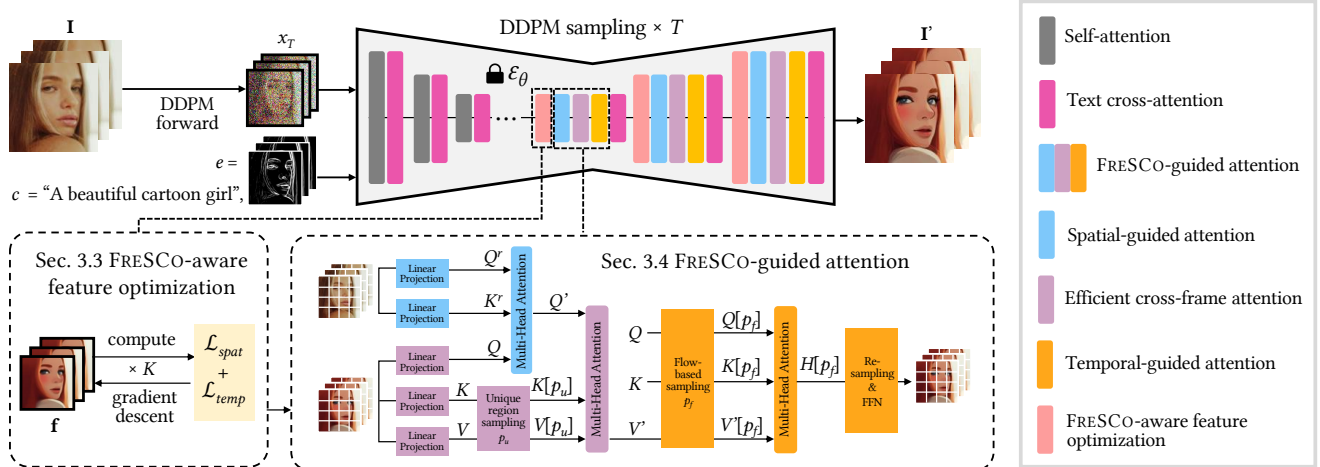
Figure 3. Framework of our zero-shot video translation guided by FRamE Spatial-temporal COrrespondence (FRESCO). A FRESCO-aware optimization is applied to the U-Net features to strengthen their temporal and spatial coherence with the input frames. We integrate FRESCO into self-attention layers, resulting in spatial-guided attention to keep spatial correspondence with the input frames, efficient cross-frame attention and temporal-guided attention to keep rough and fine temporal correspondence with the input frames, respectively.

- **Feature adaptation**. We propose a novel FRESCO-aware feature optimization approach as illustrated in Fig. 3. We design a spatial consistency loss $\mathcal{L}_{spat}$ and a temporal consistency loss $\mathcal{L}_{temp}$ to directly optimize the decoder-layer features $\mathbf{f} = \{f_i\}_{i=1}^N$ to strengthen their temporal and spatial coherence with the input frames.
- **Attention adaptation**. We replace self-attentions with FRESCO-guided attentions, comprising three components, as shown in Fig. 3. Spatial-guided attention first aggregates features based on the self-similarity of the input frame. Then, cross-frame attention is used to aggregate features across all frames. Finally, temporal-guided attention aggregates features along the same optical flow to further reinforce temporal consistency.

The proposed feature adaptation directly optimizes the feature towards high spatial and temporal coherence with **I**. Meanwhile, our attention adaptation indirectly improves coherence by imposing soft constraints on how and where to attend to valid features. We find that combining these two forms of adaptation achieves the best performance.

3.3. FRESCO-Aware Feature Optimization

The input feature $\mathbf{f} = \{f_i\}_{i=1}^N$ of each decoder layer of U-Net is updated by gradient descent through optimizing

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \mathcal{L}_{temp}(\mathbf{f}) + \mathcal{L}_{spat}(\mathbf{f}). \quad (4)$$

The updated $\hat{\mathbf{f}}$ replaces $\mathbf{f}$ for subsequent processing.

For the temporal consistency loss $\mathcal{L}_{temp}$, we would like the feature values of the corresponding positions between every two adjacent frames to be consistent,

$$\mathcal{L}_{temp}(\mathbf{f}) = \sum_i \|M_i^{i+1}(f_{i+1} - w_i^{i+1}(f_i))\|_1 \quad (5)$$

For the spatial consistency loss $\mathcal{L}_{spat}$, we use the cosine similarity in the feature space to measure the spatial correspondence of I_i. Specifically, we perform a single-step DDPM forward and backward process over I_i, and extract the U-Net decoder feature denoted as f_i^r. Since a single-step forward process adds negligible noises, f_i^r can serve as a semantic meaningful representation of I_i to calculate the semantic similarity. Then, the cosine similarity between all pairs of elements can be simply calculated as the gram matrix of the normalized feature. Let $\tilde{f}$ denote the normalized f so that each element of $\tilde{f}$ is a unit vector. We would like the gram matrix of $\tilde{f}_i$ to approach the gram matrix of $\tilde{f}_i^r$,

$$\mathcal{L}_{spat}(\mathbf{f}) = \lambda_{\text{spat}} \sum_i \|\tilde{f}_i \tilde{f}_i^\top - \tilde{f}_i^r \tilde{f}_i^{r\top}\|_2^2. \quad (6)$$

3.4. FRESCO-Guided Attention

A FRESCO-guided attention layer contains three consecutive modules: spatial-guided attention, efficient cross-frame attention and temporal-guided attention, as shown in Fig. 3. **Spatial-guided attention.** In contrast to self-attention, patches in spatial-guided attention aggregate each other based on the similarity of patches before translation rather than their own similarity. Specifically, consistent with calculating $\mathcal{L}_{spat}$ in Sec. 3.3, we perform a single-step DDPM forward and backward process over I_i, and extract its self-attention query vector Q_i^r and key vector K_i^r. Then, spatial-guided attention aggregate Q_i with

$$Q_i' = Softmax(\frac{Q_i^r K_i^{r\top}}{\lambda_s \sqrt{d}}) \cdot Q_i, \quad (7)$$

where λ_s is a scale factor and d is the query vector dimension. As shown in Fig. 4, the foreground patch will mainly
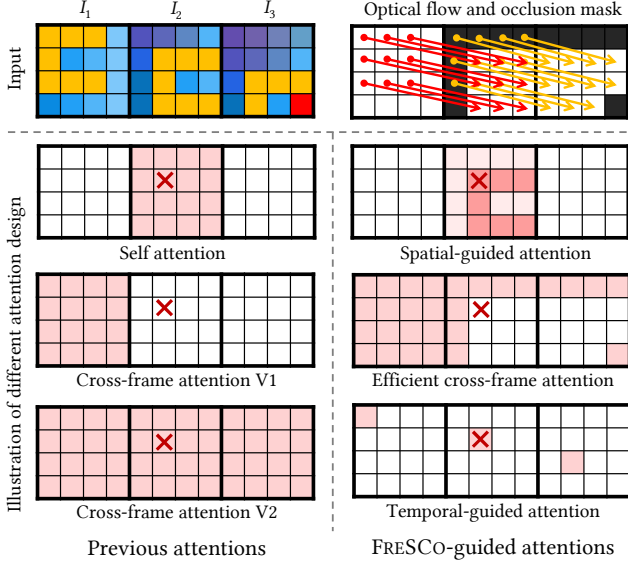
Figure 4. Illustration of attention mechanism. The patches marked with red crosses attend to the colored patches and aggregate their features. Compared to previous attentions, FRESCO-guided attention further considers intra-frame and inter-frame correspondences of the input. Spatial-guided attention aggregates intra-frame features based on the self-similarity of the input frame (darker indicates higher weights). Efficient cross-frame attention eliminates redundant patches and retains unique patches. Temporal-guided attention aggregates inter-frame features on the same flow.

aggregate features in the C-shaped foreground region, and attend less to the background region. As a result, Q' has better spatial consistency with the input frame than Q.

Efficient cross-frame attention. We replace self-attention with cross-frame attention to regularize the global style consistency. Rather than using the first frame or the previous frame as reference [4, 23] (V1, Fig. 4), which cannot handle the newly emerged objects (*e.g.*, fingers in Fig. 2(b)), or using all available frames as reference (V2, Fig. 4), which is computationally inefficient, we aim to consider all frames simultaneously but with as little redundancy as possible. Thus, we propose efficient cross-frame attentions: Except for the first frame, we only reference to the areas of each frame that were not seen in its previous frame (*i.e.*, the occlusion region). Thus, we can construct a cross-frame index p_u of all patches within the above region. Keys and values of these patches can be sampled as $K[p_u]$, $V[p_u]$. Then, cross-frame attention is applied

$$V'_i = Softmax(\frac{Q'_i(K[p_u])^\top}{\sqrt{d}}) \cdot V[p_u]. \qquad (8)$$

Temporal-guided attention. Inspired by FLATTEN [5], we use flow-based attention to regularize fine-level cross-frame consistency. We trace the same patches in different frames as in Fig. 4. For each optical flow, we build a cross-frame index p_f of all patches on this flow. In FLAT-

---

Algorithm 1 block

Algorithm 1 Keyframe selection

Input: Video $\mathbf{I} = \{I_i\}_{i=1}^M$, sample parameters $s_{\min}$, $s_{\max}$
Output: Keyframe index list Ω in ascending order
1: initialize $\Omega = [1, M]$ and $d_i = 0, \forall i \in [1, M]$
2: set $d_i = L_2(I_i, I_{i-1}), \forall i \in [s_{\min} + 1, N - s_{\min}]$
3: **while exists** i such that $\Omega[i+1] - \Omega[i] > s_{\max}$ **do**
4: $\quad \Omega.\mathtt{insert}(\hat{i}).\mathtt{sort}()$ with $\hat{i} = \arg\max_i(d_i)$
5: $\quad$ set $d_j = 0, \forall\, j \in (\hat{i} - s_{\min}, \hat{i} + s_{\min})$

TEN, each patch can only attend to patches in other frames, which is unstable when a flow contains few patches. Different from it, the temporal-guided attention has no such limit,

$$H[p_f] = Softmax(\frac{Q[p_f](K[p_f])^\top}{\lambda_t \sqrt{d}}) \cdot V'[p_f], \qquad (9)$$

where λ_t is a scale factor. And H is the final output of our FRESCO-guided attention layer.

3.5. Long Video Translation

The number of frames N that can be processed at one time is limited by GPU memory. For long video translation, we follow Rerender-A-Video [47] to perform zero-shot video translation on keyframes only and use Ebsynth [21] to interpolate non-keyframes based on translated keyframes.
Keyframe selection. Rerender-A-Video [47] uniformly samples keyframes, which is suboptimal. We propose a heuristic keyframe selection algorithm as summized in Algorithm 1. We relax the fixed sampling step to an interval $[s_{\min}, s_{\max}]$, and densely sample keyframes when motions are large (measured by L_2 distance between frames).
Keyframe translation. With over N keyframes, we split them into several N-frame batches. Each batch includes the first and last frames in the previous batch to impose inter-batch consistency, *i.e.*, keyframe indexes of the k-th batch are $\{1, (k-1)(N-2)+2, (k-1)(N-2)+3, ..., k(N-2)+2\}$. Besides, throughout the whole denoising steps, we record the latent features x'_t (Eq. (2)) of the first and last frames of each batch, and use them to replace the corresponding latent features in the next batch.

4. Experiments

Implementation details. The experiment is conducted on one NVIDIA Tesla V100 GPU. By default, we set batch size $N \in [6, 8]$ based on the input video resolution, the loss weight $\lambda_{\mathrm{spat}} = 50$, the scale factors $\lambda_s = \lambda_t = 5$. For feature optimization, we update $\mathbf{f}$ for $K = 20$ iterations with Adam optimizer and learning rate of 0.4. We find optimization mostly converges when $K = 20$ and larger K does not bring obvious gains. GMFlow [45] is used to estimate optical flows and occlusion masks. Background smoothing [23] is applied to improve temporal consistency in the background region.
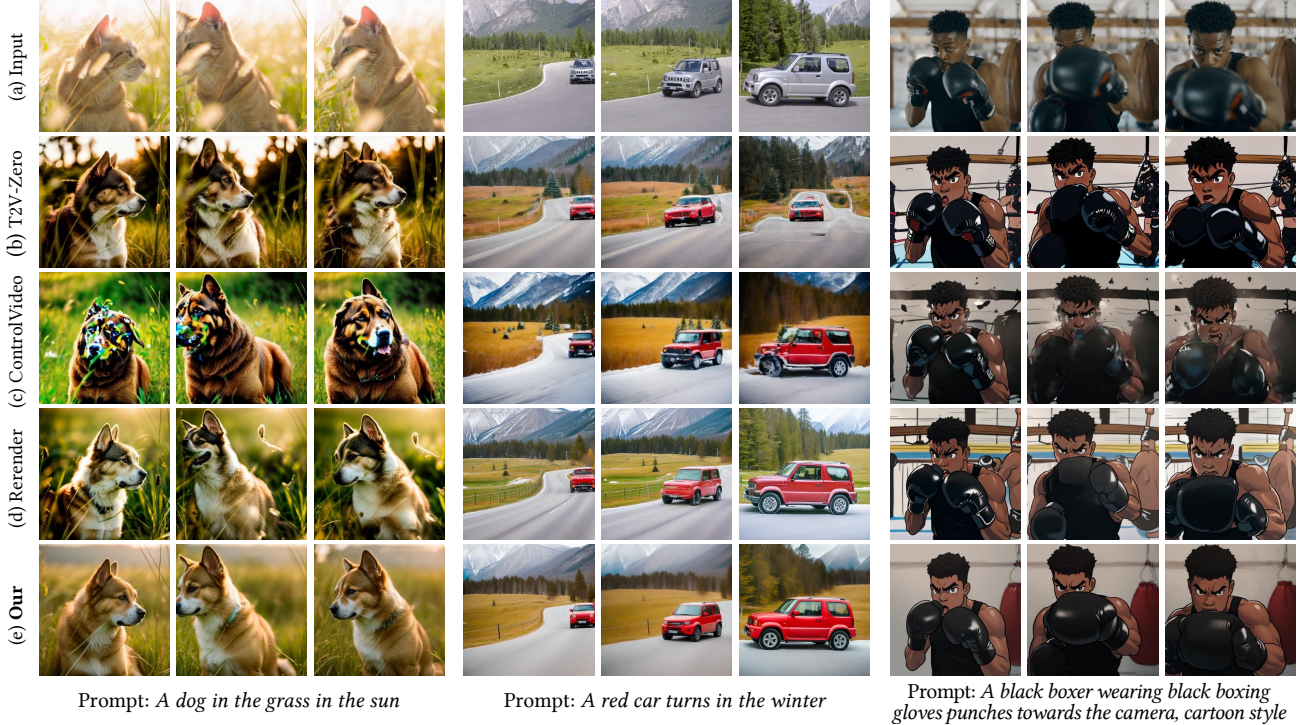
Figure 5. Visual comparison with inversion-free zero-shot video translation methods.

Prompt: *A dog in the grass in the sun*

Prompt: *A red car turns in the winter*

Prompt: *A black boxer wearing black boxing gloves punches towards the camera, cartoon style*

4.1. Comparison with State-of-the-Art Methods

We compare with three recent inversion-free zero-shot methods: Text2Video-Zero [23], ControlVideo [50], Rerender-A-Video [47]. To ensure a fair comparison, all methods employ identical settings of ControlNet, SDEdit, and LoRA. As shown in Fig. 5, all methods successfully translate videos according to the provided text prompts. However, the inversion-free methods, relying on Control-Net conditions, may experience a decline in video editing quality if the conditions are of low quality, due to issues like defocus or motion blur. For instance, ControlVideo fails to generate a plausible appearance of the dog and the boxer. Text2Video-Zero and Rerender-A-Video struggle to maintain the cat's pose and the structure of the boxer's gloves. In contrast, our method can generate consistent videos based on the proposed robust FRESCO guidance.

For quantitative evaluation, adhering to standard practices [4, 31, 47], we employ the evaluation metrics of Fram-Acc (CLIP-based frame-wise editing accuracy), Tmp-Con (CLIP-based cosine similarity between consecutive frames) and Pixel-MSE (averaged mean-squared pixel error between aligned consecutive frames). We further report Spat-Con (L_{spat} on VGG features) for spatial coherency. The results averaged across 23 videos are reported in Table 1. Notably, our method attains the best editing accuracy and temporal consistency. We further conduct a user study with 57 participants. Participants are tasked with selecting the most preferable results among the four methods. Table 1 presents

Table 1. Quantitative comparison and user preference rates.

Metric	Fram-Acc ↑	Tem-Con ↑	Pixel-MSE ↓	Spat-Con ↓	User ↑
T2V-Zero	0.918	0.965	0.038	0.0845	9.1%
ControlVideo	0.932	0.951	0.066	0.0957	2.6%
Rerender	0.955	0.969	0.016	0.0836	23.3%
Ours	**0.978**	**0.975**	**0.012**	**0.0805**	**65.0%**

Table 2. Quantitative ablation study.

Metric	baseline	w/ temp	w/ spat	w/ attn	w/ opt	full
Fram-Acc ↑	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Tem-Con ↑	0.974	0.979	0.976	0.976	0.977	**0.980**
Pixel-MSE ↓	0.032	0.015	0.020	0.016	0.019	**0.012**

the average preference rates across the 11 test videos, revealing that our method emerges as the most favored choice.

4.2. Ablation Study

To validate the contributions of different modules to the overall performance, we systematically deactivate specific modules in our framework. Figure 6 illustrates the effect of incorporating spatial and temporal correspondences. The baseline method solely uses cross-frame attention for temporal consistency. By introducing the temporal-related adaptation, we observe improvements in consistency, such as the alignment of textures and the stabilization of the sun's position across two frames. Meanwhile, the spatial-related adaptation aids in preserving the pose during translation.

In Fig. 7, we study the effect of attention adaptation and feature adaption. Clearly, each enhancement individually

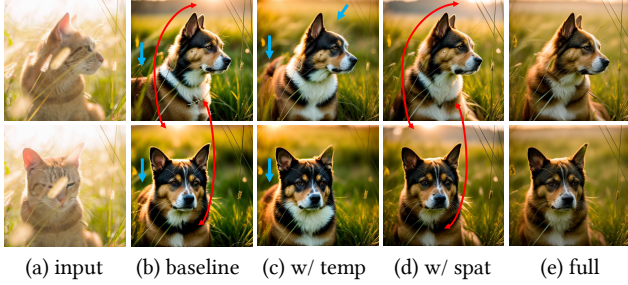(a) input   (b) baseline   (c) w/ temp   (d) w/ spat   (e) full

Figure 6. Effect of incorporating spatial and temporal correspondences. The blue arrows indicate the spatial inconsistency with the input frames. The red arrows indicate the temporal inconsistency between two output frames.
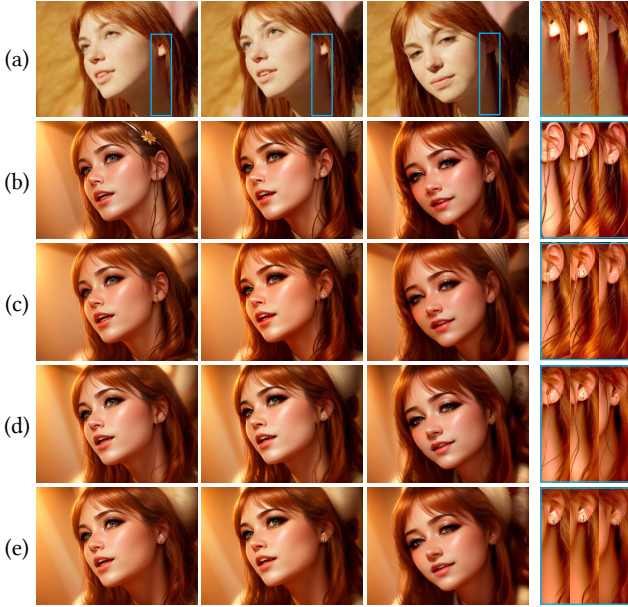

Figure 7. Effect of attention adaptation and feature adaptation. Top row: (a) Input. Other rows: Results obtained with (b) only cross-frame attention, (c) attention adaptation, (d) feature adaptation, (e) both attention and feature adaptations, respectively. The blue region is enlarged with its contrast enhanced on the right for better comparison. Prompt: A beautiful woman in CG style.

improves temporal consistency to a certain extent, but neither achieves perfection. Only the combination of the two completely eliminates the inconsistency observed in hair strands, which is quantitatively verified by the Pixel-MSE scores of 0.037, 0.021, 0.018, 0.015 for Fig. 7(b)-(e), respectively. Regarding attention adaptation, we further delve into temporal-guided attention and spatial-guided attention. The strength of the constraints they impose is determined by λ_t and λ_s, respectively. As shown in Figs. 8-9, an increase in λ_t effectively enhances consistency between two transformed frames in the background region, while an increase in λ_s boosts pose consistency between the transformed cat and the original cat. Beyond spatial-guided attention, our spatial consistency loss also plays an important role, as validated in Fig. 10. In this example, rapid motion and blur
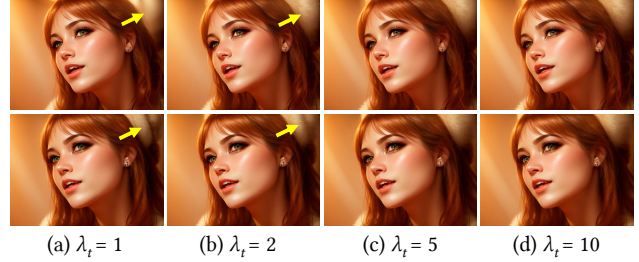

(a) $\lambda_t = 1$ (b) $\lambda_t = 2$ (c) $\lambda_t = 5$ (d) $\lambda_t = 10$

Figure 8. Effect of λ_t. Quantitatively, the Pixel-MSE scores are (a) 0.016, (b) 0.014, (c) 0.013, (d) 0.012. The yellow arrows indicate the inconsistency between the two frames.
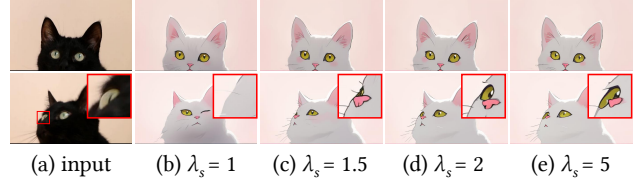

(a) input   (b) $\lambda_s = 1$ (c) $\lambda_s = 1.5$ (d) $\lambda_s = 2$ (e) $\lambda_s = 5$

Figure 9. Effect of λ_s. The region in the red box is enlarged and shown in the top right for better comparison. Prompt: A cartoon white cat in pink background.

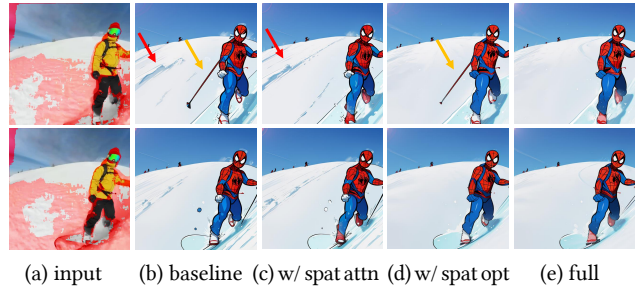(a) input   (b) baseline   (c) w/ spat attn   (d) w/ spat opt   (e) full

Figure 10. Effect of incorporating spatial correspondence. (a) Input covered with red occlusion mask. (b)-(d) Our spatial-guided attention and spatial consistency loss help reduce the inconsistency in ski poles (yellow arrows) and snow textures (red arrows), respectively. Prompt: A cartoon Spiderman is skiing.

make optical flow hard to predict, leading to a large occlusion region. Spatial correspondence guidance is particularly crucial to constrain the rendering in this region. Clearly, each adaptation makes a distinct contribution, such as eliminating the unwanted ski pole and inconsistent snow textures. Combining the two yields the most coherent results, as quantitatively verified by the Pixel-MSE scores of 0.031, 0.028, 0.025, 0.024 for Fig. 10(b)-(e), respectively.

Table 2 provides a quantitative evaluation of the impact of each module. In alignment with the visual results, it is evident that each module contributes to the overall enhancement of temporal consistency. Notably, the combination of all adaptations yields the best performance.

Figure 11 ablates the proposed efficient cross-frame attention. As with Rerender-A-Video in Fig. 2(b), sequential frame-by-frame translation is vulnerable to new appearing objects. Our cross-frame attention allows attention to all unique objects within the batched frames, which is not only efficient but also more robust, as demonstrated in Fig. 12.

(a) input   (b) Cross-frame attention V1

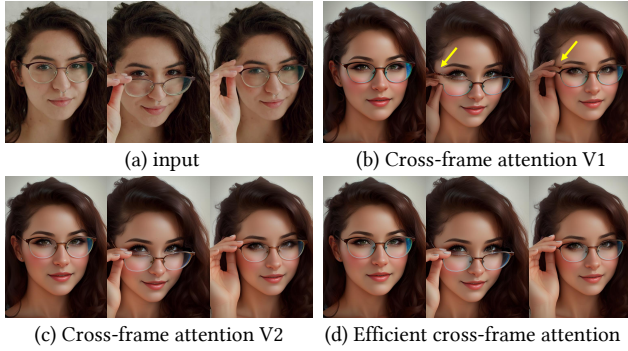(c) Cross-frame attention V2   (d) Efficient cross-frame attention

Figure 11. Effect of efficient cross-frame attention. (a) Input. (b) Cross-frame attention V1 attends to the previous frame only, thus failing to synthesize the newly appearing fingers. (d) The efficient cross-frame attention achieves the same performance as (c) cross-frame attention V2, but reduces the region that needs to be attended to by 41.6% in this example. Prompt: A beautiful woman holding her glasses in CG style.

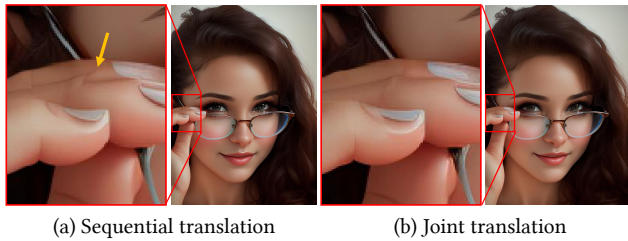(a) Sequential translation   (b) Joint translation

Figure 12. Effect of joint multi-frame translation. Sequential translation relies on the previous frame alone. Joint translation uses all frames in a batch to guide each other, thus achieving accurate finger structures by referencing to the third frame in Fig. 11

(a) optimize *f before attention* (b) optimize *f* after attention   (c) optimize $\hat{x}_0$

Figure 13. Diffusion features to optimize.

FRESCO uses diffusion features before the attention layers for optimization. Since U-Net is trained to predict noise, features after attention layers (near output layer) are noisy, leading to failure optimization (Fig. 13(b)). Meanwhile, the four-channel $\hat{x}_0'$ (Eq. (3)) is highly compact, which is not suitable for warping or interpolation. Optimizing $\hat{x}_0'$ results in severe blurs and over-saturation artifacts (Fig. 13(c)).

4.3. More Results

Long video translation. Figure 1 presents examples of long video translation. A 16-second video comprising 400 frames are processed, where 32 frames are selected as keyframes for diffusion-based translation and the remaining 368 non-keyframes are interpolated. Thank to our FRESCO guidance to generate coherent keyframes, the non-keyframes exhibit coherent interpolation as in Fig. 14.
Video colorization. Our method can be applied to video colorization. As shown in Fig. 15, by combining the L chan-

keyframe #40   non-keyframe #42   non-keyframe #43   non-keyframe #45

non-keyframe #47   non-keyframe #48   non-keyframe #50   keyframe # 52

Figure 14. Long video generation by interpolating non-keyframes based on the translated keyframes.

Figure 15. Video colorization. Prompt: A blue seal on the beach.

nel from the input and the AB channel from the translated video, we can colorize the input without altering its content.

4.4. Limitation and Future Work

In terms of limitations, first, Rerender-A-Video directly aligns frames at the pixel level, which outperforms our method given high-quality optical flow. We would like to explore an adaptive combination of these two methods in the future to harness the advantages of each. Second, by enforcing spatial correspondence consistency with the input video, our method does not support large shape deformations and significant appearance changes. Large deformation makes it challenging to use the optical flow of the original video as a reliable prior for natural motion. This limitation is inherent in zero-shot models. A potential future direction is to incorporate learned motion priors [13].

5. Conclusion

This paper presents a zero-shot framework to adapt image diffusion models for video translation. We demonstrate the vital role of preserving intra-frame spatial correspondence, in conjunction with inter-frame temporal correspondence, which is less explored in prior zero-shot methods. Our comprehensive experiments validate the effectiveness of our method in translating high-quality and coherent videos. The proposed FRESCO constraint exhibits high compatibility with existing image diffusion techniques, suggesting its potential application in other text-guided video editing tasks, such as video super-resolution and colorization.

References

[1] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proc. Int'l Conf. Computer Vision*, 2023. 3

[4] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proc. Int'l Conf. Computer Vision*, pages 23206–23217, 2023. 1, 3, 5, 6

[5] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. FLATTEN: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023. 1, 2, 3, 5

[6] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proc. Int'l Conf. Computer Vision*, pages 7346–7356, 2023. 1, 3

[7] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccedit: Creative and controllable video editing via diffusion models. *arXiv preprint arXiv:2309.16496*, 2023. 3

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Proc. Int'l Conf. Learning Representations*, 2022. 3

[9] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics*, 42(4):1–13, 2023. 3

[10] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proc. Int'l Conf. Computer Vision*, 2023. 3

[11] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *Proc. Int'l Conf. Learning Representations*, 2024. 1, 3

[12] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. TaleCrafter: Interactive story visualization with multiple characters. In *ACM SIGGRAPH Asia Conference Proceedings*, 2023. 3

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3, 8

[14] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. SVDiff: Compact parameter space for diffusion fine-tuning. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2023. 3

[15] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 3

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *Proc. Int'l Conf. Learning Representations*, 2022. 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 2, 3

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 3

[19] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *Proc. Int'l Conf. Learning Representations*, 2021. 1, 3

[20] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023. 3

[21] Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. Stylizing video by example. *ACM Transactions on Graphics*, 38(4):1–11, 2019. 5

[22] Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. *arXiv preprint arXiv:2310.01107*, 2023. 3

[23] 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 *Proc. Int'l Conf. Computer Vision*, 2023. 1, 2, 3, 5, 6

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2023. 3

[25] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-P2P: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 1

[26] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. VideoFusion: Decomposed diffusion models for high-quality video generation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 3

[27] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jia-jun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *Proc. Int'l Conf. Learning Representations*, 2021. 3

[28] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3

[29] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proc. IEEE Int'l Conf. Machine Learning*, pages 16784–16804, 2022. 2

[30] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH Conference Proceedings*, pages 1–11, 2023. 3

[31] 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 *Proc. Int'l Conf. Computer Vision*, 2023. 1, 3, 6

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. IEEE Int'l Conf. Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2

[35] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, pages 36479–36494, 2022. 1, 2

[37] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-A-Video: Single video editing with object-aware consistency. *arXiv preprint arXiv:2303.07945*, 2023. 1, 3

[38] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-A-Video: Text-to-video generation without text-video data. In *Proc. Int'l Conf. Learning Representations*, 2023. 1, 3

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proc. Int'l Conf. Learning Representations*, 2021. 3

[40] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 3

[41] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 1

[42] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 3

[43] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proc. Int'l Conf. Computer Vision*, 2023. 3

[44] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. In *Proc. Int'l Conf. Computer Vision*, pages 7623–7633, 2023. 1, 2, 3

[45] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. GMFlow: Learning optical flow via global matching. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 8121–8130, 2022. 5

[46] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking "text" out of text-to-image diffusion models. *arXiv preprint arXiv:2305.16223*, 2023. 3

[47] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *ACM SIGGRAPH Asia Conference Proceedings*, 2023. 1, 2, 3, 5, 6

[48] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. Int'l Conf. Computer Vision*, pages 3836–3847, 2023. 1, 3

[50] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. ControlVideo: Training-free controllable text-to-video generation. In *Proc. Int'l Conf. Learning Representations*, 2024. 3, 6

[51] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 3