Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation

Shuai Yang Yifan Zhou Ziwei Liu Chen Change Loy
S-Lab, Nanyang Technological University
{shuai.yang, yifan.zhou, ziwei.liu, ccloy}@ntu.edu.sg

Figure 1. We propose a novel zero-shot text-guided video-to-video translation framework based on pre-trained image diffusion model.

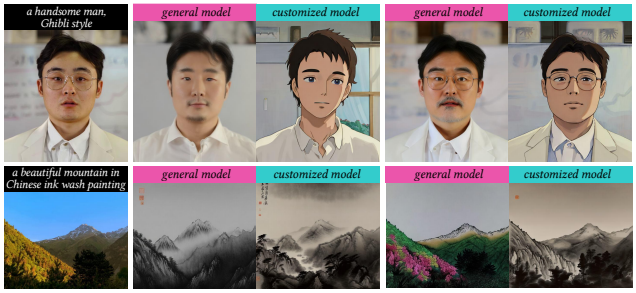
Abstract

Large text-to-image diffusion models have exhibited impressive proficiency in generating high-quality images. However, when applying these models to video domain, ensuring temporal consistency across video frames remains a formidable challenge. This paper proposes a novel zero-shot text-guided video-to-video translation framework to adapt image models to videos. The framework includes two parts: key frame translation and full video translation. The first part uses an adapted diffusion model to generate key frames, with hierarchical cross-frame constraints applied to enforce coherence in shapes, textures and colors. The second part propagates the key frames to other frames with temporal-aware patch matching and frame blending. Our framework achieves global style and local texture temporal consistency at a low cost (without re-training or optimization). The adaptation is compatible with existing image diffusion techniques, allowing our framework to take advantage of them, such as customizing a specific subject with LoRA, and introducing extra spatial guidance with ControlNet. Extensive experimental results demonstrate the effectiveness of our proposed framework over existing methods in rendering high-quality and temporally-coherent videos. Project page https://anonymous-31415926.github.io/. Code will be released.

1. Introduction

Recent text-to-image diffusion models such as DALLE-2 [26], Imagen [30], Stable Diffusion [28] demonstrate exceptional ability in generating diverse and high-quality images guided by natural language. Based on it, a multitude of image editing methods have emerged, including model fine-tuning for customized object generation [29], image-to-image translation [20], image inpainting [1], and object editing [10]. These applications allow users to synthesize and edit images effortlessly, using natural language within a unified diffusion framework, greatly improving creation efficiency. As video content surges in popularity on social media platforms, the demand for more streamlined video creation tools has concurrently risen. Yet, a critical challenge remains: the direct application of existing image diffusion models to videos leads to severe flickering issues.

Researchers have recently turned to text-to-video diffusion models and proposed three solutions. The first solution involves training a video model on large-scale video data [11], which requires significant computing resources.



(a) input (b) Translation results w/o ControlNet (c) Translation results w/ ControlNet

Figure 2. Customized model and ControlNet generate high-quality results with better consistency with both prompt and content. Our method is designed to be compatible with these existing image diffusion techniques, and thus can take advantage of them to strike a good balance between the style (prompt) and the content.

Additionally, the re-designed video model is incompatible with existing off-the-shelf image models. The second solution is to fine-tune image models on a single video [36], which is less efficient for long videos. Overfitting to a single video may also degrade the performance of the original models. The third solution involves zero-shot methods [17] that require no training. During the diffusion sampling process, cross-frame constraints are imposed on the latent features for temporal consistency. The zero-shot strategy requires fewer computing resources and is mostly compatible with existing image models, showing promising potential. However, current cross-frame constraints are limited to global styles and are unable to preserve low-level consistency, *e.g.*, the overall style may be consistent, but the local structures and textures may still flicker.

Achieving successful application of image diffusion models to the video domain is a challenging task. It requires **1)** Temporal consistency: cross-frame constraints for low-level consistency; **2)** Zero-shot: no training or fine-tuning required; **3)** Flexibility: compatible with off-the-shelf image models for customized generation. As mentioned above, image models can be customized by fine-tuning on specific objects to capture the target style more precisely than general models. Figure 2 shows two examples. To take advantage of it, in this paper, we employ zero-shot strategy for model compatibility and aim to further solve the key issue of this strategy in maintaining low-level temporal consistency.

To achieve this goal, we propose novel hierarchical cross-frame constraints for pre-trained image models to produce coherent video frames. Our key idea is to use optical flow to apply dense cross-frame constraints, with the previous rendered frame serving as a low-level reference for the current frame and the first rendered frame acting as an anchor to regulate the rendering process to prevent deviations from the initial appearance. Hierarchical cross-frame constraints are realized at different stages of diffusion sampling. In addition to global style consistency, our method enforces consistency in shapes, textures and colors at early, middle and late stages, respectively. This innovative and lightweight modification achieves both global and local temporal consistency. Figure 1 presents our coherent video translation results over off-the-shelf image models customized for six unique styles.

Based on the insight, this paper introduces a novel zero-shot framework for text-guided video-to-video translation, consisting of two parts: key frame translation and full video translation. In the first part, we adapt pre-trained image diffusion models with hierarchical cross-frame constraints for generating key frames. In the second part, we propagate the rendered key frames to other frames using temporal-aware patch matching and frame blending. The diffusion-based generation is excellent at content creation, but its multi-step sampling process is inefficient. The patch-based propagation, on the other hand, can efficiently infer pixel-level coherent frames but is not capable of creating new content. By combining these two parts, our framework strikes a balance between quality and efficiency. To summarize, our main contributions are as follows:

- A novel zero-shot framework for text-guided video-to-video translation, which achieves both global and local temporal consistency, requires no training, and is compatible with pre-trained image diffusion models.
- Hierarchical cross-frame consistency constraints to enforce temporal consistency in shapes, textures and colors, which adapt image diffusion models to videos.
- Hybrid diffusion-based generation and patch-based propagation to strike a balance between quality and efficiency.

2. Related Work

2.1. Text Driven Image Generation

Generating images with descriptive sentences is intuitive and flexible. Early attempts explore GAN [38–40, 42] to synthesize realistic images. With the powerful expressivity of Transformer [34], autoregressive models [5, 7, 27] are proposed to model image pixels as a sequence with autoregressive dependency between each pixel. DALL-E [27] and CogView [5] train an autoregressive transformer on image and text tokens. Make-A-Scene [7] further considers segmentation masks as condition.

Recent studies focus on diffusion models [12] for text-to-image generation, where images are synthesized via a gradual denoising process. DALLE-2 [26] and Imagen [30] introduce pretrained large language models [24, 25] as text encoder to better align the image with text, and cascade diffusion models for high resolution image generation. GLIDE [22] introduces classifier-free guidance to improve text conditioning. Instead of applying denoising in the image space, Latent Diffusion Models [28] uses the low-resolution latent space of VQ-GAN [6] to improve the efficiency. We refer to [4] for a thorough survey.

In addition to diffusion models for general images, customized models are studied. Textual Inversion [8] and DreamBooth [29] learn special tokens to capture novel concepts and generate related images given a small number of example images. LoRA [14] accelerates the fine-tuning large models by learning low-rank weight matrices added to existing weights. ControlNet [41] fine-tunes a new control path to provide pixel-level conditions such as edge maps and pose, enabling fine-grained image generation. Our method does not alter the pre-trained model, thus is orthogonal to these existing techniques. This empowers our method to leverage DreamBooth and LoRA for better customized video translation and to use ControlNet for temporal-consistent structure guidance as in Fig. 2.

2.2. Video Editing with Diffusion Models

For text-to-video generation, Video Diffusion Model [13] proposes to extend the 2D U-Net in image model to a factorized space-time UNet. Imagen Video [11] scales up the Video Diffusion Model with a cascade of spatial and temporal video super-resolution models. Make-A-Video [32] leverages video data in an unsupervised manner to learn the movement to drive the image model. Although promising, the above methods need large-scale video data for training.

Tune-A-Video [36] instead inflates an image diffusion model into a video model with cross-frame attention, and fine-tunes it on a single video to generate videos with related motion. Based on it, Edit-A-Video [31], Video-P2P [19] and vid2vid-zero [35] utilize Null-Text Inversion [21] for precise inversion to preserve the unedited region. However, these models need fine-tuning of the pre-trained model or optimization over the input video, which is less efficient.

Recent developments have seen the introduction of zero-shot methods that, by design, operate without any training phase. Thus, these methods are naturally compatible with pre-trained diffusion variants like InstructPix2Pix [2] or ControlNet to accept more flexible conditions like depth and edges. Based on the editing masks detected by Prompt2Prompt [10] to indicate the channel and spatial region to preserve, FateZero [23] blends the attention features before and after editing. Text2Video-Zero [17] translates the latent to directly simulate motions and Pix2Video [3] matches the latent of the current frame to that of the previous frame. All the above methods largely rely on cross-frame attention and early-step latent fusion to improve temporal consistency. However, as we will show later, these strategies predominantly cater to high-level styles and shapes, and being less effective in maintaining cross-frame consistency at the level of texture and detail. In contrast to these approaches, our method proposes a novel pixel-aware cross-frame latent fusion, which non-trivially achieves pixel-level temporal consistency.

3. Preliminary: Diffusion Models

Stable Diffusion Stable Diffusion is a latent diffusion model operating in the latent space of an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, where $\mathcal{E}$ and are the encoder and decoder, respectively. Specifically, for an image I with its latent feature $x_0 = \mathcal{E}(I)$, the diffusion forward process iteratively add noises to the latent

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{1}$$

where $t = 1, ..., T$ is the time step, $q(x_t|x_{t-1})$ is the conditional density of x_t given g_{t-1}, and α_t is hyperparameters. Alternatively, we can directly sample x_t at any time step from x_0 with,

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

Then in the diffusion backward process, a U-Net ϵ_θ is trained to predict the noise of the latent to iteratively recover x_0 from x_T. Given a large T, x_0 will be completely destroyed in the forward process so that x_T approximates a standard Gaussian distribution. Therefore, ϵ_θ correspondingly learns to infer valid x_0 from random Gaussian noises. Once trained, we can sample x_{t-1} based on x_t with a deterministic DDIM sampling [33]:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\hat{x}_{t\to 0}}_{\text{predicted } x_0} + \underbrace{\sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t, t, c_p)}_{\text{direction pointing to } x_{t-1}}, \tag{3}$$

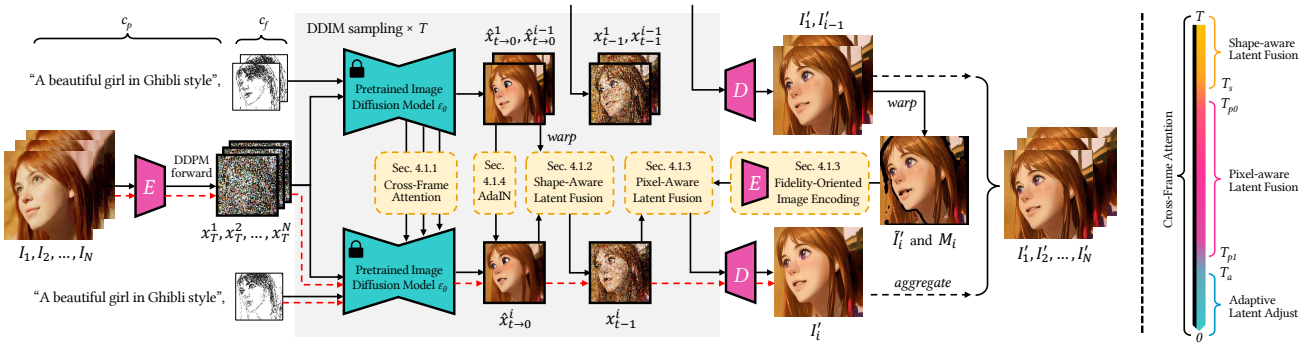
where $\hat{x}_{t\to 0}$ is the predicted x_0 at time step t,

$$\hat{x}_{t\to 0} = (x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t, c_p))/\sqrt{\alpha_t}, \tag{4}$$

and $\epsilon_\theta(x_t, t, c_p)$ is the predicted noise of x_t based on the time step t and the text prompt condition c_p.

During inference, we can sample a valid x_0 from the standard Guassian noise $x_T = z_T, z_T \sim \mathcal{N}(0, \mathbf{I})$ with DDIM sampling, and decode x_0 to the final generated image $I' = \mathcal{D}(x_0)$.

ControlNet Although flexible, natural language has limited spatial control over the output. To improve spatial controllability, [41] introduce a side path called ControlNet to Stable Diffusion to accept extra conditions like edges, depth and human pose. Let c_f be the extra condition, the noise prediction of U-Net with ControlNet becomes $\epsilon_\theta(x_t, t, c_p, c_f)$. Compared to InstructPix2Pix, ControlNet is orthogonal to customized Stable Diffusion models. To build a general zero-shot V2V framework, we use ControlNet to provide structure guidance from the input video to improve temporal consistency.



(a) Diffusion-based key frame translation with hierarchical cross-frame consistency constraints (b) Sampling pipeline

Figure 3. Framework of the proposed zero-shot text-guided video translation. (a) We adapt the pre-trained image diffusion model (Stable Diffusion + ControlNet) with hierarchical cross-frame constraints to render coherent frames. The red dotted lines denote the sampling process of the original image diffusion model. The black lines denote our adapted process for video translation. (b) We apply different constraints at different sampling steps.

4. Zero-Shot Text-Guided Video Translation

Given a video with N frames $\{I_i\}_{i=0}^N$, our goal is to render it into a new video $\{I'_i\}_{i=0}^N$ in another artistic expression specified by text prompts and/or off-the-shelf customized Stable Diffusion models. Our framework consists of two parts: Key Frame Translation (Sec. 4.1) and Full Video Translation (Sec. 4.2). In the first part, we introduce hierarchical cross-frame constraints into the pre-trained diffusion models to render coherent key frames. Then the second part propagates the rendered key frames to other frames with temporal-aware matching and blending.

4.1. Key Frame Translation

Figure 3 illustrates the T-step sampling pipeline for the key frame translation. Following SDEdit [20], the pipeline begins with $x_T = \sqrt{\bar{\alpha}_T}x_0 + (1-\bar{\alpha}_T)z_T, z_T \sim \mathcal{N}(0, \mathbf{I})$, the noisy latent code of the input video frame rather than the pure Gaussian noise. It enables users to determine how much detail of the input frame is preserved in the output by adjusting T, *i.e.*, smaller T retain more detail. Then, during sampling each frame, we use the first frame as anchor frame and its previous frame to constrain global style consistency and local temporal consistency.

Specifically, cross-frame attention [36] is applied to all sampling steps for global style consistency (Sec. 4.1.1). In addition, in early steps, we fuse the latent feature with the aligned latent feature of previous frame to achieve rough shape alignments (Sec. 4.1.2). Then in mid steps, we use the latent feature with the encoded warped anchor and previous outputs to realize fine texture alignments (Sec. 4.1.3). Finally, in late steps, we adjust the latent feature distribution for color consistency (Sec. 4.1.4). For simplicity, we will use $\{I_i\}_{i=0}^N$ to refer to the key frames in this section.

4.1.1 Style-aware cross-frame attention

Similar to other zero-shot video editing methods [3, 17], we replace self-attention layers in the U-Net with cross-frame attention layers to regularize the global style of I'_i to match that of I'_1 and I'_{i-1}. In Stable Diffusion, each self-attention layer receives the latent feature v_i (for simplicity we omit the time step t) of I_i, and linearly projects v_i into query, key and value Q, K, V to produce the output by $\textit{Self\_Attn}(Q,K,V) = \textit{Softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V$ with

$$Q = W^Q v_i, K = W^K v_i, V = W^V v_i, \tag{5}$$

where W^Q, W^K, W^V are pre-trained matrices for feature projection. Cross-frame attention, by comparison, uses the key K' and value V' from other frames (we use the first and previous frames), *i.e.*, $\textit{CrossFrame\_Attn}(Q,K',V') = \textit{Softmax}(\frac{QK'^T}{\sqrt{d}}) \cdot V'$ with

$$Q = W^Q v_i, K' = W^K [v_1; v_{i-1}], V' = W^V [v_1; v_{i-1}]. \tag{6}$$

Intuitively, self-attention can be thought as patch matching and voting within a single frame, while cross-frame attention seeks similar patches and fuses the corresponding patches from other frames, meaning the style of I'_i will inherit that of I'_1 and I'_{i-1}.

4.1.2 Shape-aware cross-frame latent fusion

Cross-frame attention is limited to global style. To constrain the cross-frame local shape and texture consistency, we use optical flow to warp and fuse the latent features. Let w^i_j and M^i_j denote the optical flow and occlusion mask from I_j to I_i, respectively. Let x^i_t be the latent feature for I'_i at time step t. We update the predicted $\hat{x}_{t\to 0}$ in Eq. (3) by

$$\hat{x}^i_{t\to 0} = M^i_j \cdot \hat{x}^i_{t\to 0} + (1 - M^i_j) \cdot w^i_j(\hat{x}^j_{t\to 0}). \tag{7}$$

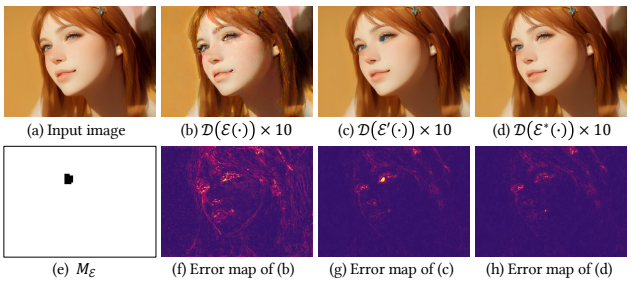


Figure 4. Fidelity-oriented image encoding.

w and M are downsampled to match the resolution of x (we omit the downsampling operation for simplicity in this paper). For the reference frame I_j, we experimentally find that the anchor frame ($j = 0$) provides better guidance than the previous frame ($j = i - 1$). We observe that interpolating elements in the latent space can lead to blurring and shape distortion in the late steps. Therefore, we limit the fusion to only early steps for rough shape guidance.

4.1.3 Pixel-aware cross-frame latent fusion

To constrain the low-level texture features in mid steps, instead warping the latent feature, we can alternatively warp previous frames and encode them back to the latent space for fusion. However, the lossy autoencoder introduces distortions and color bias that easily accumulate along the frame sequence. Figure 4(b) shows an example of the distorted result after encoding and decoding 10 times. [1] solved this problem by fine-tuning the decoder's weights to fit each image, which is impractical for long videos. To efficiently solve this problem, we propose a novel fidelity-oriented zero-shot image encoding method.

Fidelity-oriented image encoding Our key insight is the observation that the amount of information lost each time in the iterative auto-encoding process is consistent. Therefore, we can predict the information loss for compensation. Specifically, for arbitrary image I, we encode and decode it twice, obtaining $x_0^r = \mathcal{E}(I), I_r = \mathcal{D}(x_0^r)$ and $x_0^{rr} = \mathcal{E}(I_r), I_{rr} = \mathcal{D}(x_0^{rr})$. We assume the loss from the target lossless x_0 to x_0^r is linear to that from x_0^r to x_0^{rr}. Then we define the encoding $\mathcal{E}'$ with compensation as

$$\mathcal{E}'(I) := x_0^r + \lambda_\mathcal{E}(x_0^r - x_0^{rr}), \tag{8}$$

where we find the linear coefficient $\lambda_\mathcal{E} = 1$ works well. We further add a mask $M_\mathcal{E}$ to prevent the possible artifacts introduced by compensation (*e.g.*, blue artifact near the eyes in Fig. 4(c)). $M_\mathcal{E}$ indicates where the error between I and $\mathcal{D}(\mathcal{E}'(I))$ is under a pre-defined threshold. Then, our novel fidelity-oriented image encoding $\mathcal{E}^*$ takes the form of

$$\mathcal{E}^*(I) := x_0^r + M_\mathcal{E} \cdot \lambda_\mathcal{E}(x_0^r - x_0^{rr}). \tag{9}$$

As shown in Fig. 4(d), our method preserves image information well even after encoding and decoding 10 times.

Structure-guided inpainting For pixel-level coherence, we warp the anchor frame I'_0 and the previous frame I'_{i-1} to the i-th frame and overlay them on a rough rendered frame $\bar{I}'_i$ obtained without the pixel-aware cross-frame latent fusion as

$$M_0^i \cdot \big(M_{i-1}^i \cdot \bar{I}'_i + (1 - M_{i-1}^i) \cdot w_{i-1}^i(I'_{i-1})\big) + \big(1 - M_0^i\big) \cdot w_0^i(I'_0) \tag{10}$$

The resulting fused frame $\tilde{I}'_i$ provides pixel reference for the sampling of I'_i, *i.e.*, we would like I'_i to match $\tilde{I}'_i$ outside the mask area $M_i = M_0^i \cap M_{i-1}^i$ and to match the structure guidance from ControlNet inside M_i. We formulate it as a structure-guided inpainting task and follow [1] to inpaint x_{t-1}^i in Eq. (3) as

$$x_{t-1}^i = M_i \cdot x_{t-1}^i + (1 - M_i) \cdot \tilde{x}_{t-1}^i, \tag{11}$$

where $\tilde{x}_{t-1}^i$ is the sampled x_{t-1} from $x_0 = \mathcal{E}^*(\tilde{I}'_i)$ based on Eq. (2).

4.1.4 Color-aware adaptive latent adjustment

Finally, we apply AdaIN [15] to $\hat{x}_{t\to 0}^i$ to match its channel-wise mean and variance to $\hat{x}_{t\to 0}^1$ in the late steps. It can further keep the color style coherent throughout the whole key frames.

4.2. Full Video Translation

For frames with similar content, existing video frame interpolation algorithms are capable of generating plausible results by propagating the rendered key frames to their neighbors efficiently. However, compared to diffusion models, frame interpolation cannot create new content. To balance between quality and efficiency, we apply our adapted diffusion model to the key frames only. Then we adopt a patch-based frame interpolation algorithm [16] to render the remaining frames based on the coherent key frames.

Specifically, we sample the key frames from the video uniformly for every K frame. The key frames are denoted by $I_0, I_K, I_{2K}, ...$ and are rendered to $I'_0, I'_K, I'_{2K}, ...$ by our adapted diffusion model. Taking the first K frames for example, we interpolate $I'_1 \sim I'_{K-1}$ with I'_0 and $'_K$. First, we use a patch-based propagation to propagate I'_0 and obtain the intermediate frames $I'^0_1 \sim I'^0_{K-1}$. Likewise, we propagate I'_K and obtain $I'^K_1 \sim I'^K_{K-1}$. Then we adopt the temporal-aware blending to blend I'^0_i and I'^K_i to get the final result I'_i. Please refer to [16] for the details.

4.2.1 Patch-based propagation

The patch-based propagation aims to find a dense correspondences between the key frame and its neighbor frames and use the resulting correspondences map to warp the coherent key frame. We adopt a guided path-matching algorithm [16] with color, positional, edge, and temporal guidance for dense correspondence prediction.

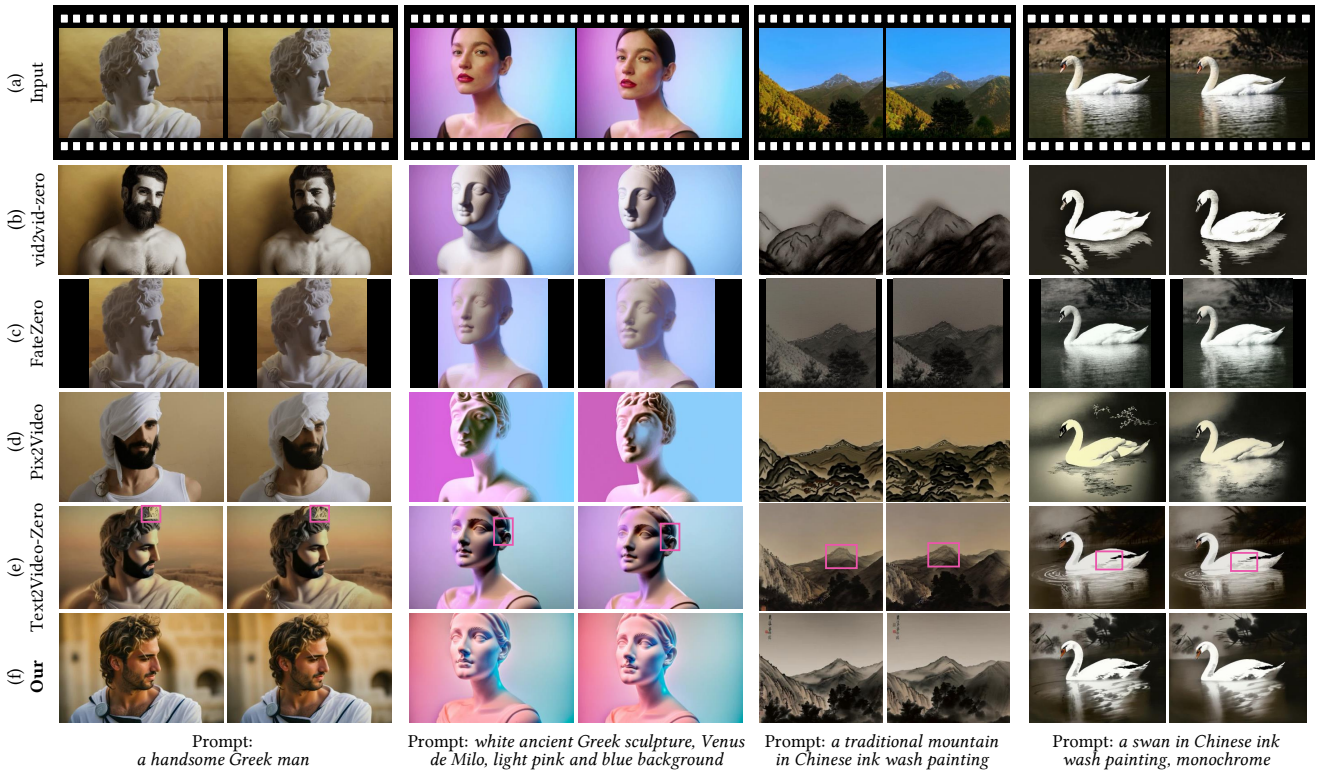


Figure 5. Visual comparison with zero-shot video translation methods. The red box indicates the inconsistent region.

4.2.2 Temporal-aware blending

The main idea of the blending of $I_i'^0$ and $I_i'^K$ is to combine them with a pixel selection mask M_i based on the patch matching error. Specifically, M_i comes from a minimum error mask $\hat{M}_i$ denoting the region where the patch matching error is lower on $I_i'^0$ than $I_i'^K$, *i.e.*, $\hat{M}_i(p) = 1$ if $E_i^0(p) < E_i^K(p)$ where $E_i^0(p)$ is the patch match error between $I_i'^0$ and I_0' at pixel p. To preserve the temporal consistency, M_0 starts from a all-one matrix, and during the update, the previous mask M_{i-1} is warped to an initial $\bar{M}_i$ with optical flow w_{i-1}. Then only the pixels satisfy $\hat{M}_i(p) = 0$ are allowed to update $\bar{M}_i(p)$ to obtain M_i. This measure can prevent a pixel selecting the next key frame from turning back to the previous key frame. The blended frame serves as the histogram reference for contrast-preserving blending [9] over $I_i'^0$ and $I_i'^K$ to obtain the final result. Different from [16], we do not apply Poisson fusion, which we find sometimes blurs the textures and causes flickers around the key frame.

5. Experimental Results

5.1. Implementation Details

The experiment is conducted on one NVIDIA Tesla V100 GPU. We employ the fine-tuned and LoRA models based on Stable Diffusion 1.5 from https://civitai.com/. We use Stable Diffusion originally uses $T_{max} = 1000$ steps. For the sampling pipeline in Fig. 3(b), by default, we set $T_s = 0.1T_{max}$, $T_{p0} = 0.5T_{max}$, $T_{p1} = 0.8T_{max}$ and $T_a = 0.8T_{max}$. We tune T for each video. ControlNet [41] is used to provide structure guidance in terms of edges, with the control weight tuned for each video. We use GMFlow [37] for optical flow estimation. For full video translation, by default, we sample key frames for every $K = 10$ frames. The testing videos are from https://www.pexels.com/ and https://pixabay.com/, with their short side resized to 512.

In terms of running time for 512×512 videos, key frame and non-key frame translations take about 14.23s and 1.49s per frame, respectively. Overall, a full video translation takes about $(14.23 + 1.49(K-1))/K = 1.49 + 12.74/K$s per frame.

We will release our code upon publication of the paper.

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

We compare with 4 recent zero-shot methods: vid2vid-zero [35], FateZero [23], Pix2Video [3], Text2Video-Zero [17] on key frame translation with $K = 5$. The official code of the first three methods does not support ControlNet, and when loading customized models, we find they fail to generate plausible results, *e.g.*, vid2vid-zero will generate frames totally different from the input. Therefore, only Text2Video-Zero and our method use the customized

Table 1. User preference rates in terms of the balance between the content and the prompt, temporal consistency and overall quality.

Method	Balance	Temporal	Overall
vid2vid-zero	3.8%	4.3%	3.8%
FateZero	6.0%	9.2%	4.9%
Pix2Video	10.3%	5.5%	4.9%
Text2Video-Zero	15.2%	13.1%	15.2%
Ours	**64.7%**	**67.9%**	**71.2%**

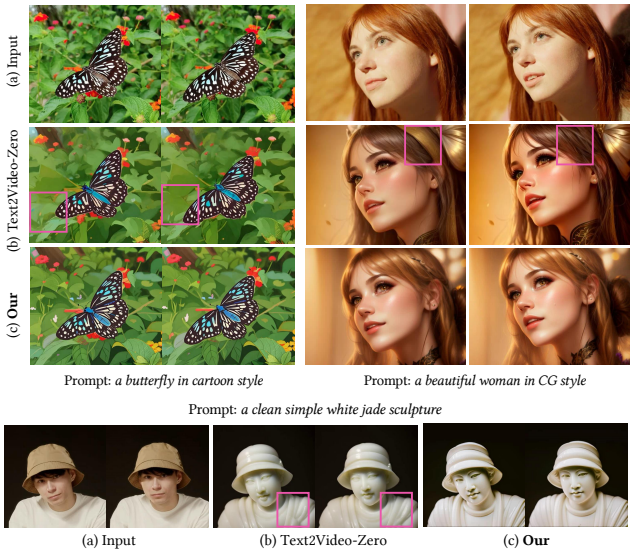


Figure 6. Visual comparison with Text2Video-Zero. Text2Video-Zero and our method use the same customized model and ControlNet for a fair comparison. Our method outperforms Text2Video-Zero in terms of local texture temporal consistency. The red box indicates the inconsistent region.

model with ControlNet. Figure 5 and Figure 6 present the visual results. FateZero successfully reconstructs the input frame but fails to adjust it to match the prompt. On the other hand, vid2vid-zero and Pix2Video excessively modify the input frame, leading to significant shape distortion and discontinuity across frames. While each frame generated by Text2Video-Zero exhibits high quality, they lack coherence in local textures. Finally, our proposed method demonstrates clear superiority in terms of output quality, content and prompt matching and temporal consistency.

As ground truth data is not available, using conventional metrics like FVD is not feasible. To evaluate the quality quantitatively, we conduct a user study with 23 participants. The participants are asked to select the best results among the five methods based on three criteria: 1) how well the result balance between the prompt and the input frame, 2) the temporal consistency of the result, and 3) the overall quality of the video translation. Table 1 presents the average preference rates across 8 testing videos, and our method achieves the highest rates in all three metrics.

5.3. Ablation Study

Hierarchical cross-frame consistency constraints Figure 7 compares the results with and without different cross-frame consistency constraints. We demonstrate the efficacy of our approach on a video containing simple translational motion in the first half and complex 3D rotation transformations in the latter half. To better evaluate the temporal consistency, we encourage readers to watch the videos on the project webpage. The cross-frame attention ensures consistency in global style, while the adaptive latent adjustment in Sec. 4.1.4 maintains the same hair color as the first frame, or the hair color will follow the input frame to turn dark. Note that the adaptive latent adjustment is optional to allow users to decide which color to follow. The above two global constraints cannot capture local movement. The shape-aware latent fusion (SA fusion) in Sec. 4.1.2 addresses this by translating the latent features to translate the neck ring, but cannot maintain pixel-level consistency for complex motion. Only the proposed pixel-aware latent fusion (PA fusion) can coherently render local details such as hair styles and acne.

We provide additional examples in Figs. 8-9 to demonstrate the effectiveness of PA fusion. While ControlNet can guide the structure well, the inherent randomness introduced by noise addition and denoising makes it difficult to maintain coherence in local textures, resulting in missing elements and altered details. The proposed PA fusion restores these details by utilizing the corresponding pixel information from previous frames. Moreover, such consistency between key frames can effectively reduce the ghosting artifacts in interpolated non key frames.

Fidelity-oriented image encoding We present a detailed analysis of our fidelity-oriented image encoding in Figs. 10-12, in addition to Fig. 4. Two Stable Diffusion's officially released autoencoders, the fine-tuned f8-ft-MSE VAE and the original more lossy kl-f8 VAE, are used for testing our method. The fine-tuned VAE introduces artifacts and the original VAE results in great color bias as in Fig. 10(b). Our proposed fidelity-oriented image encoding effectively alleviates these issues. For quantitative evaluation, we report the MSE between the input image and the reconstructed result after multiple encoding and decoding in Fig. 11, using the first 1,000 images of the MS-COCO [18] validation set. The results are consistent with the visual observations: our proposed method significantly reduces error accumulation compared to raw encoding methods. Finally, we validate our encoding method in the video translation process in Fig. 12(b)(c), where we use only the previous frame without the anchor frame in Eq. (10) to better visualize error accumulation. Our method mostly reduces the loss of details and color bias caused by lossy encoding. Besides,

Figure 7. Effect of the proposed hierarchical cross-frame constraints. (a) Input frames #1, #55, #94. (b) Image diffusion model renders each frame independently. (c) Cross frame attention keeps the overall style consistent. (d) AdaIN preserves the hair color. (e) Shape-aware latent fusion keeps the overall movement of the objects coherent. (f) Pixel-aware latent fusion achieves pixel-level temporal consistency.

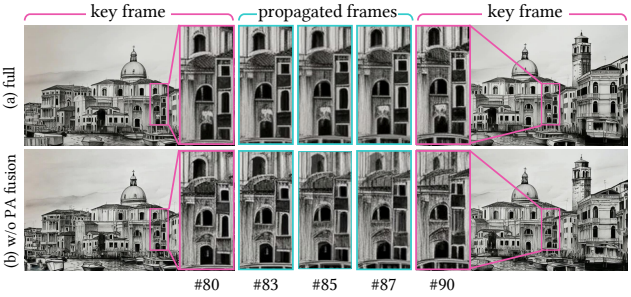


Figure 8. Effect of the pixel-aware latent fusion on frame propagation. The proposed pixel-aware latent fusion helps generate consistent key frames. Without it, the pixel level inconsistency between key frames leads to ghosting artifacts on the non-key frames during the frame blending.

our pipeline includes an anchor frame and adaptive latent adjustment to further regulate the translation, as shown in Fig. 12(d), where no obvious errors are observed.

5.4. More Results

Flexible structure and color control The proposed pipeline allows flexible control over content preservation through the initialization of x_T. Rather than setting x_T to a Gaussian noise (Fig. 13(b)), we use a noisy latent version of the input frame to better preserve details (Fig. 13(c)). Users can adjust the value of T to balance content and prompt. Moreover, if the input frame introduces unwanted color bias (*e.g.*, blue sky in Chinese ink painting), a color correction option is provided: the input frame is adjusted to match the color histogram of the frame generated by $x_T = z_T$ (Fig. 13(b)). With the adjusted frame as input (bottom row of Fig. 13(a)), the rendered results (bottom row of Figs. 13(c)-(f)) better match the color indicated by the prompt.

Figure 9. Effect of the pixel-aware latent fusion. Prompts (from top to bottom): '*Arcane style, a handsome man*', '*Loving Vincent, hiking, grass*', '*Disco Elysium, street view*'. Local regions are enlarged and shown in the top right.

Applications Figure 14 shows some applications of our method. With prompts 'a cute cat/fox/hamster/rabbit', we can perform text-guided editing to translate a dog into other kinds of pets in Fig. 14(a). By using customized modes for generating cartoons or photos, we can achieve non-photorealistic and photorealistic rendering in Fig. 14(b) and

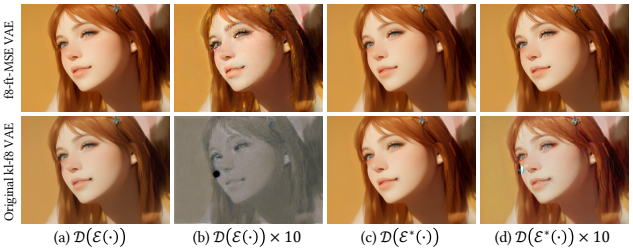


Figure 10. The fidelity-oriented image encoding on two VAEs.

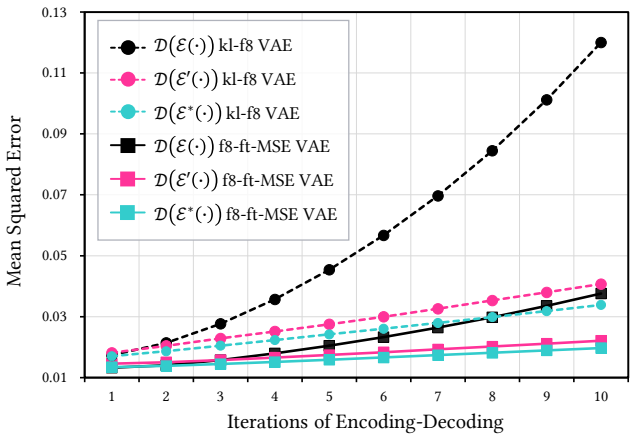


Figure 11. Quantitative evaluation of image encoding schemes.

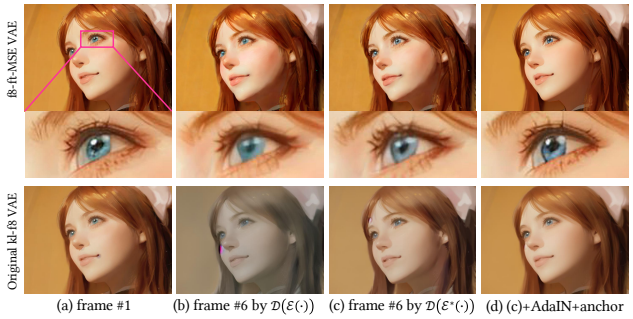


Figure 12. Different constraints to prevent error accumulation.

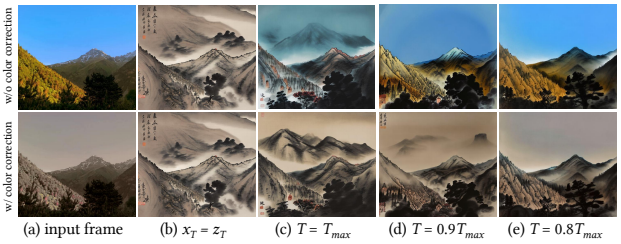


Figure 13. Effect of the initialization of x_T. Prompt: a traditional mountain in Chinese ink wash painting. The proposed framework enables flexible content and color control by adjusting T and color correction.

Figs. 14(c)(d), respectively. In Fig. 15, we present our synthesized dynamic virtual characters of novels and manga, based on a real human video and a prompt to describe the appearance. Additional results are shown in Fig. 16.

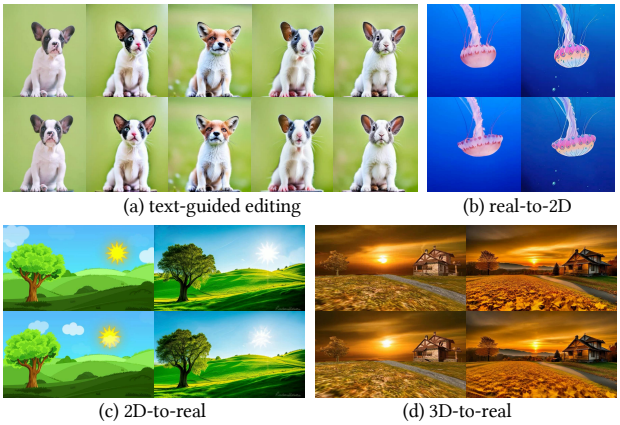


Figure 14. Applications of the proposed method.

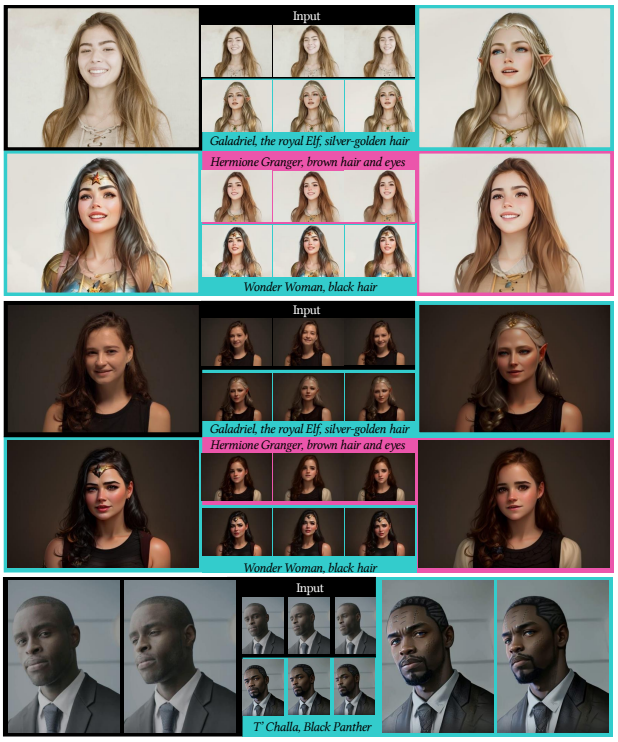


Figure 15. Applications: text-guided virtual character generation. Results are generated with a single image diffusion model.

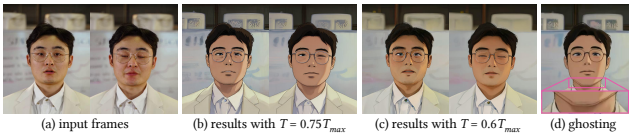
5.5. Limitations

Figures 17-19 illustrate typical failure cases of our method. First, our method relies on optical flow and therefore, inaccurate optical flow can lead to artifacts. In Fig. 17, our method can only preserve the embroidery if the cross-frame correspondence is available. Otherwise, the proposed PA fusion will have no effect. Second, our method assumes the optical flow remains unchanged before and after translation, which may not hold true for significant appearance changes as in Fig. 18(b), where the resulting movement may be wrong. Although setting a smaller T can address this issue, it may compromise the desired styles. Meanwhile, the mismatches of the optical flow mean the mismatches in the

Figure 16. Applications: video stylization. Thanks to the compatible design, our method can use off-the-shelf pre-trained image models customized for different styles to accurately stylize videos.

(a) frame #1 (b) Results w/ pixel-aware fusion (c) Results w/o pixel-aware fusion

Figure 17. Limitation: failure optical flow due to large motions. Our method is not suitable for processing videos where it is difficult to estimate the optical flow.



(a) input frames (b) results with $T = 0.75T_{max}$ (c) results with $T = 0.6T_{max}$ (d) ghosting

Figure 18. Limitation: trade-off between content and prompt.

translated key frames, which may lead to ghosting artifacts (Fig. 18(d)) after temporal-aware blending. Also, we find that small details and subtle motions like accessories and eye movement cannot be well preserved during the translation. Lastly, we uniformly sample the key frames, which may not optimal. Ideally, the key frames should contain all unique objects; otherwise, the propagation cannot create unseen content such as the hand in Fig. 19(b). One potential solution is user-interactive translation, where users can manually assign new key frames based on the previous results.

6. Conclusion

This paper presents a zero-shot framework to adapt image diffusion models for video translation. Our method utilizes hierarchical cross-frame constraints to enforce temporal consistency in both global style and low-level textures, leveraging the key optical flow. The compatibility with existing image diffusion techniques indicates that our idea

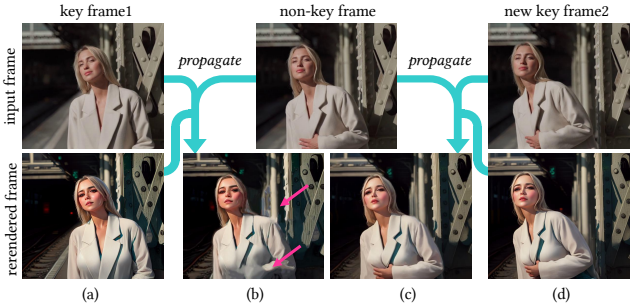


Figure 19. Limitation: failed propagation w/o good key frames.

might be applied to other text-guided video editing tasks, such as video super-resolution and inpainting. Additionally, our proposed fidelity-oriented image encoding could benefit existing diffusion-based methods. We believe that our approach can facilitate the creation of high-quality and temporally-coherent videos and inspire further research in this field.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 1, 5

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 3

[3] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. *arXiv preprint arXiv:2303.12688*, 2023. 3, 4, 6

[4] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[5] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 19822–19835, 2021. 2

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 2

[7] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Proc. European Conf. Computer Vision*, pages 89–106. Springer, 2022. 2

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[9] Eric Heitz and Fabrice Neyret. High-performance by-example noise using a histogram-preserving blending oper-

ator. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 1(2):1–25, 2018. 6

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 3

[11] 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

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020. 2

[13] Jonathan Ho, Tim Salimans, Alexey A Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022. 3

[14] 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. 3

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. Int'l Conf. Computer Vision*, pages 1510–1519, 2017. 5

[16] Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sỳkora. Stylizing video by example. *ACM Transactions on Graphics*, 38(4):1–11, 2019. 5, 6

[17] 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. *arXiv preprint arXiv:2303.13439*, 2023. 2, 3, 4, 6

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. European Conf. Computer Vision*, pages 740–755. Springer, 2014. 7

[19] 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. 3

[20] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun 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. 1, 4

[21] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 3

[22] 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

[23] 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. *arXiv preprint arXiv:2303.09535*, 2023. 3, 6

[24] 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

[25] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2

[26] 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

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proc. IEEE Int'l Conf. Machine Learning*, pages 8821–8831, 2021. 2

[28] 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

[29] 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. *arXiv preprint arXiv:2208.12242*, 2022. 1, 3

[30] 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*, volume 35, pages 36479–36494, 2022. 1, 2

[31] 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. 3

[32] 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. 3

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

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 2

[35] 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. 3, 6

[36] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, 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. *arXiv preprint arXiv:2212.11565*, 2022. 2, 3, 4

[37] 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. 6

[38] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 1316–1324, 2018. 2

[39] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 833–842, 2021. 2

[40] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[41] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3, 6

[42] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 5802–5810, 2019. 2