Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators

Levon Khachatryan[1*] Andranik Movsisyan[1*] Vahram Tadevosyan[1*] Roberto Henschel[1*]
Zhangyang Wang[1,2] Shant Navasardyan[1] Humphrey Shi[1,3]
[1]Picsart AI Resarch (PAIR) [2]UT Austin [3]SHI Labs @ Georgia Tech, Oregon & UIUC
https://github.com/Picsart-AI-Research/Text2Video-Zero

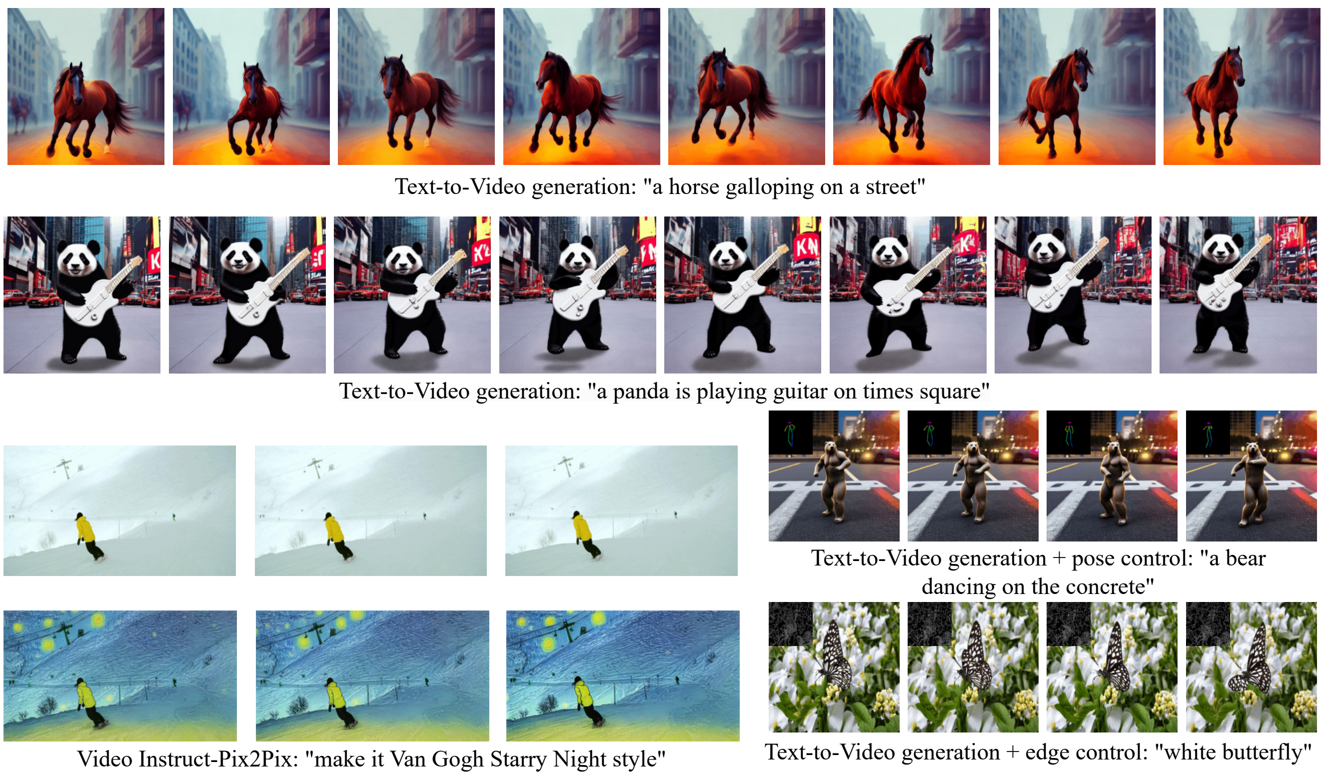


Figure 1: Our method Text2Video-Zero enables zero-shot video generation using (i) a textual prompt (see rows 1, 2), (ii) a prompt combined with guidance from poses or edges (see lower right), and (iii) Video Instruct-Pix2Pix, *i.e.*, instruction-guided video editing (see lower left). Results are temporally consistent and follow closely the guidance and textual prompts.

Abstract

*Recent text-to-video generation approaches rely on computationally heavy training and require large-scale video datasets. In this paper, we introduce a new task, zero-shot text-to-video generation, and propose a low-cost approach (without any training or optimization) by leveraging the power of existing text-to-image synthesis methods (*e.g. *Stable Diffusion), making them suitable for the video domain. Our key modifications include (i) enriching the latent codes of the generated frames with motion dynamics to keep the global scene and the background time consistent; and (ii) reprogramming frame-level self-attention using a new cross-frame attention of each frame on the first frame, to preserve the context, appearance, and identity of the foreground object. Experiments show that this leads to low overhead, yet high-quality and remarkably consistent video generation. Moreover, our approach is not limited to text-to-video synthesis but is also applicable to other tasks such as conditional and content-specialized video generation, and Video Instruct-Pix2Pix,* i.e., *instruction-guided*

[*]Equal contribution.

video editing. As experiments show, our method performs comparably or sometimes better than recent approaches, despite not being trained on additional video data. Our code is publicly available at: https://github.com/Picsart-AI-Research/Text2Video-Zero.

1. Introduction

In recent years, generative AI has attracted enormous attention in the computer vision community. With the advent of diffusion models [34, 12, 35, 36], it has become tremendously popular and successful to generate high-quality images from textual prompts, also called *text-to-image synthesis* [26, 29, 32, 7, 44]. Recent works [14, 33, 11, 42, 5, 21] attempt to extend the success to text-to-video generation and editing tasks, by reusing text-to-image diffusion models in the video domain. While such approaches yield promising outcomes, most of them require substantial training with a massive amount of labeled data which can be costly and unaffordable for many users. With the aim of making video generation cheaper, Tune-A-Video [42] introduces a mechanism that can adopt Stable Diffusion (SD) model [29] for the video domain. The training effort is drastically reduced to tuning one video. While that is much more efficient than previous approaches, it still requires an optimization process. In addition, the generation abilities of Tune-A-Video are limited to text-guided video editing applications; video synthesis from scratch, however, remains out of its reach.

In this paper, we take one step forward in studying the novel problem of *zero-shot, "training-free" text-to-video synthesis*, which is the task of generating videos from textual prompts without requiring any optimization or fine-tuning. A key concept of our approach is to modify a pre-trained text-to-image model (*e.g*., Stable Diffusion), enriching it with temporally consistent generation. By building upon already trained text-to-image models, our method takes advantage of their excellent image generation quality and enhances their applicability to the video domain without performing additional training. To enforce temporal consistency, we present two innovative and lightweight modifications: (1) we first enrich the latent codes of generated frames with motion information to keep the global scene and the background time consistent; (2) we then use cross-frame attention of each frame on the first frame to preserve the context, appearance, and identity of the foreground object throughout the entire sequence. Our experiments show that these simple modifications lead to high-quality and time-consistent video generations (see Fig. 1 and further results in the appendix). Despite the fact that other works train on large-scale video data, our method achieves similar or sometimes even better performance (see Figures 8, 9 and appendix Figures 18, 25, 26). Furthermore, our method is not limited to text-to-video synthesis but is also applicable to conditional (see Figures 5,6 and appendix Figures 19, 21, 22, 23) and specialized video generation (see Fig. 7), and instruction-guided video editing, which we refer as *Video Instruct-Pix2Pix* motivated by Instruct-Pix2Pix [2] (see Fig. 9 and appendix Figures 24, 25, 26).

Our contributions are summarized as three-folds:

- A new problem setting of zero-shot text-to-video synthesis, aiming at making text-guided video generation and editing "freely affordable". We use only a pre-trained text-to-image diffusion model without any further fine-tuning or optimization.
- Two novel post-hoc techniques to enforce temporally consistent generation, via encoding motion dynamics in the latent codes, and reprogramming each frame's self-attention using a new cross-frame attention.
- A broad variety of applications that demonstrate our method's effectiveness, including conditional and specialized video generation, and *Video Instruct-Pix2Pix* *i.e*., video editing by textual instructions.

2. Related Work

2.1. Text-to-Image Generation

Early approaches to text-to-image synthesis relied on methods such as template-based generation [19] and feature matching [28]. However, these methods were limited in their ability to generate realistic and diverse images.

Following the success of GANs [8], several other deep learning-based methods were proposed for text-to-image synthesis. These include StackGAN [46], AttnGAN [43], and MirrorGAN [24], which further improve image quality and diversity by introducing novel architectures and attention mechanisms.

Later, with the advancement of transformers [38], new approaches emerged for text-to-image synthesis. Being a 12-billion-parameter transformer model, Dall-E [27] introduces a two-stage training process: First, it generates image tokens, which later are combined with text tokens for joint training of an autoregressive model. Later Parti [45] proposed a method to generate content-rich images with multiple objects. Make-a-Scene [7] enables a control mechanism by segmentation masks for text-to-image generation.

Current approaches build upon diffusion models, thereby taking text-to-image synthesis quality to the next level. GLIDE [23] improved Dall-E by adding classifier-free guidance [13]. Later, Dall-E 2 [26] utilizes the contrastive model CLIP [25]. By means of diffusion processes, (i) a mapping from CLIP text encodings to image encodings, and (ii) a CLIP decoder is obtained. LDM / SD [29] applies a diffusion model on lower-resolution encoded signals of VQ-GAN [6], showing competitive quality with a significant gain in speed and efficiency. Imagen [32] shows incredible performance in text-to-image synthesis by utilizing large language models for text processing. Versatile Diffusion [44] further unifies text-to-image, image-to-text and variations in a single multi-flow diffusion model. Specialisation of text-to-image models to desired styles can be obtained efficiently via few-shot tuning, *e.g*. using DreamBoth [31] or Specialist Diffusion [18], which employs text-to-image customized data augmentations.

Because of their great image quality, it is desired to exploit text-to-image models for video generation. However, applying diffusion models in the video domain is not straightforward, especially due to their probabilistic generation procedure, making it difficult to ensure temporal consistency. As we show in our ablation experiments in the appendix (see Fig. 14), our modifications are crucial for temporal consistency in terms of both global scene and background motion, and for the preservation of the foreground object identity.

2.2. Text-to-Video Generation

Text-to-video synthesis is a relatively new research direction. Existing approaches try to leverage autoregressive transformers and diffusion processes for the generation. NUWA [41] introduces a 3D transformer encoder-decoder framework and supports both text-to-image and text-to-video generation. Phenaki [39] introduces a bidirectional masked transformer with a causal attention mechanism that allows the generation of arbitrary-long videos from text prompt sequences. CogVideo [15] extends the text-to-image model CogView 2 [4] by tuning it using a multi-frame-rate hierarchical training strategy to better align text and video clips. Video Diffusion Models (VDM) [14] naturally extend text-to-image diffusion models and train jointly on image and video data. Imagen Video [11] constructs a cascade of video diffusion models and utilizes spatial and temporal super-resolution models to generate high-resolution time-consistent videos. Make-A-Video [33] builds upon a text-to-image synthesis model and leverages video data in an unsupervised manner. Gen-1 [5] extends SD and proposes a structure and content-guided video editing method based on visual or textual descriptions of desired outputs. Tune-A-Video [42] proposes a new task of one-shot video generation by extending and tuning SD on a single reference video.

Unlike the methods mentioned above, our approach is completely training-free, does not require massive computing power or dozens of GPUs, which makes the video generation process affordable for everyone. In this respect, Tune-a-Video [42] comes closest to our work, as it reduces the necessary computations to tuning on only one video. However, it still requires an optimization process and its generating ability is heavily restricted by the reference video.

3. Method

We start this section with a brief introduction of diffusion models, particularly Stable Diffusion (SD) [29]. Then we introduce the problem formulation of zero-shot text-to-video synthesis, followed by a subsection presenting our approach. After that, to show the universality of our method, we use it in combination with ControlNet [47] and DreamBooth [31] diffusion models for generating conditional and specialized videos. Later we demonstrate the power of our approach with the application of instruction-guided video editing, namely, Video Instruct-Pix2Pix.

3.1. Stable Diffusion

SD is a diffusion model operating in the latent space of an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, namely VQ-GAN [6] or VQ-VAE [37], where $\mathcal{E}$ and $\mathcal{D}$ are the corresponding encoder and decoder, respectively. More precisely, if $x_0 \in \mathbb{R}^{h\times w\times c}$ is the latent tensor of an input image Im given by the autoencoder, i.e. $x_0 = \mathcal{E}(\mathrm{Im})$, diffusion forward process iteratively adds Gaussian noise to the signal x_0:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),\ t = 1,..,T \quad (1)$$

where $q(x_t|x_{t-1})$ is the conditional density of x_t given x_{t-1}, and $\{\beta_t\}_{t=1}^T$ are hyperparameters. T is chosen to be as large that the forward process completely destroys the initial signal x_0 resulting in $x_T \sim \mathcal{N}(0, I)$. The goal of SD is then to learn a backward process

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)) \quad (2)$$

for $t = T,\ldots,1$, which allows to generate a valid signal x_0 from the standard Gaussian noise x_T. To get the final image generated from x_T it remains to pass x_0 to the decoder of the initially chosen autoencoder: $\mathrm{Im} = \mathcal{D}(x_0)$.

After learning the abovementioned backward diffusion process (see DDPM [12]) one can apply a deterministic sampling process, called DDIM [35]:

$$\begin{aligned} x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^t(x_t)}{\sqrt{\alpha_t}}\right) + \\ \sqrt{1-\alpha_{t-1}}\epsilon_\theta^t(x_t), \quad t = T,\ldots,1, \end{aligned} \quad (3)$$

where $\alpha_t = \prod_{i=1}^t (1-\beta_i)$ and

$$\epsilon_\theta^t(x_t) = \frac{\sqrt{1-\alpha_t}}{\beta_t}x_t + \frac{(1-\beta_t)(1-\alpha_t)}{\beta_t}\mu_\theta(x_t,t). \quad (4)$$

To get a text-to-image synthesis framework, SD guides the diffusion processes with a textual prompt τ. Particularly for DDIM sampling, we get:

$$\begin{aligned} x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^t(x_t,\tau)}{\sqrt{\alpha_t}}\right) + \\ \sqrt{1-\alpha_{t-1}}\epsilon_\theta^t(x_t,\tau), \quad t = T,\ldots,1. \end{aligned} \quad (5)$$

It is worth noting that in SD, the function $\epsilon_\theta^t(x_t, \tau)$ is modeled as a neural network with a UNet-like [30] architecture composed of convolutional and (self- and cross-) attentional blocks. x_T is called the latent code of the signal x_0 and there is a method [3] to apply a deterministic forward process to reconstruct the latent code x_T given a signal x_0. This method is known as DDIM inversion. Sometimes for simplicity, we will call $x_t, t = 1, \dots, T$ also the *latent codes* of the initial signal x_0.

3.2. Zero-Shot Text-to-Video Problem Formulation

Existing text-to-video synthesis methods require either costly training on a large-scale (ranging from $1M$ to $15M$ data-points) text-video paired data [41, 15, 39, 11, 14, 5] or tuning on a reference video [42]. To make video generation cheaper and easier, we propose a new problem: zero-shot text-to-video synthesis. Formally, given a text description τ and a positive integer $m \in \mathbb{N}$, the goal is to design a function $\mathcal{F}$ that outputs video frames $\mathcal{V} \in \mathbb{R}^{m \times H \times W \times 3}$ (for predefined resolution $H \times W$) that exhibit temporal consistency. To determine the function $\mathcal{F}$, no training or fine-tuning must be performed on a video dataset.

Our problem formulation provides a new paradigm for text-to-video. Noticeably, a zero-shot text-to-video method naturally benefits from quality improvements of text-to-image models.

3.3. Method

In this paper, we approach the zero-shot text-to-video task by exploiting the text-to-image synthesis power of Stable Diffusion (SD). As we need to generate videos instead of images, SD should operate on sequences of latent codes. The naïve approach is to independently sample m latent codes from standard Gaussian distribution $x_T^1, \dots, x_T^m \sim \mathcal{N}(0, I)$ and apply DDIM sampling to obtain the corresponding tensors x_0^k for $k = 1, \dots, m$, followed by decoding to obtain the generated video sequence $\{\mathcal{D}(x_0^k)\}_{k=1}^m \in \mathbb{R}^{m \times H \times W \times 3}$. However, this leads to completely random generation of images sharing only the semantics described by τ but neither object appearance nor motion coherence (see appendix Fig. 14, first row).

To address this issue, we propose to (i) introduce motion dynamics between the latent codes $x_T^1, \dots, x_T^m$ to keep the global scene time consistent and (ii) use cross-frame attention mechanism to preserve the appearance and the identity of the foreground object. Each of the components of our method are described below in detail. The overview of our method can be found in Fig. 2.

Note, to simplify notation, we will denote the entire sequence of latent codes by $x_T^{1:m} = [x_T^1, \dots, x_T^m]$.

Algorithm 1 Motion dynamics in latent codes

Require: $\Delta t \geq 0, m \in \mathbb{N}, \lambda > 0, \delta = (\delta_x, \delta_y) \in \mathbb{R}^2$, Stable Diffusion (SD)

$x_T^1 \sim \mathcal{N}(0, I)$ ▷ *random sample the first latent code*
$x_{T'}^1 \leftarrow$ DDIM_Backward$(x_T^1, \Delta t, SD)$ ▷ *perform Δt backward steps by SD*
for all $k = 2, 3, \dots, m$ **do**
  $\delta^k \leftarrow \lambda \cdot (k-1)\delta$ ▷ *computing global translation vectors*
  $W_k \leftarrow$ Warping by δ^k ▷ *defining warping functions*
  $\tilde{x}_{T'}^k \leftarrow W_k(x_{T'}^1)$
  $x_T^k \leftarrow$ DDPM_Forward$(\tilde{x}_{T'}^k, \Delta t)$ ▷ *DDPM forward for more motion freedom*
return $x_T^{1:m}$

3.3.1 Motion Dynamics in Latent Codes

Instead of sampling the latent codes $x_T^{1:m}$ randomly and independently from the standard Gaussian distribution, we *construct* them by performing the following steps (see also Alg. 1 and Fig. 2).

1. Randomly sample the latent code of the first frame: $x_T^1 \sim \mathcal{N}(0, I)$.
2. Perform $\Delta t \geq 0$ DDIM backward steps on the latent code x_T^1 by using the SD model and get the corresponding latent $x_{T'}^1$, where $T' = T - \Delta t$.
3. Define a direction $\delta = (\delta_x, \delta_y) \in \mathbb{R}^2$ for the global scene and camera motion. By default δ can be the main diagonal direction $\delta_x = \delta_y = 1$.
4. For each frame $k = 1, 2, \dots, m$ we want to generate, compute the global translation vector $\delta^k = \lambda \cdot (k-1)\delta$, where λ is a hyperparameter controlling the amount of the global motion.
5. Apply the constructed motion (translation) flow $\delta^{1:m}$ to $x_{T'}^1$, denote the resulting sequence by $\tilde{x}_{T'}^{1:m}$:
$$\tilde{x}_{T'}^k = W_k(x_{T'}^1) \text{ for } k = 1, 2, \dots, m, \tag{6}$$
where $W_k(x_{T'}^1)$ is the warping operation for translation by the vector δ^k.
6. Perform Δt DDPM forward steps on each of the latents $\tilde{x}_{T'}^{2:m}$ and get the corresponding latent codes $x_T^{2:m}$.

Then we take the sequence $x_T^{1:m}$ as the starting point of the backward (video) diffusion process. As a result, the latent codes generated with our proposed motion dynamics lead to better temporal consistency of the global scene as well as the background, see in the appendix, Sect. 8.1. Yet, the initial latent codes are not constraining enough to describe particular colors, identities or shapes, thus still

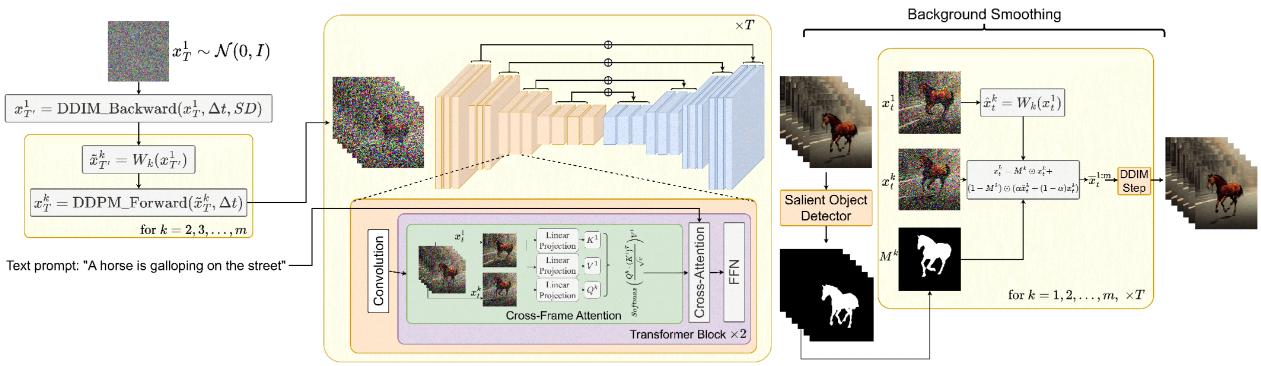


Figure 2: Method overview: Starting from a randomly sampled latent code x_T^1, we apply Δt DDIM backward steps to obtain $x_{T'}^1$ using a pre-trained Stable Diffusion model (SD). A specified motion field results for each frame k in a warping function W_k that turns $x_{T'}^1$ to $x_{T'}^k$. By enhancing the latent codes with motion dynamics, we determine the global scene and camera motion and achieve temporal consistency in the background and the global scene. A subsequent DDPM forward application delivers latent codes x_T^k for $k = 1, \ldots, m$. By using the (probabilistic) DDPM method, a greater degree of freedom is achieved with respect to the motion of objects (see appendix Sec. 8.1). Finally, the latent codes are passed to our modified SD model using the proposed cross-frame attention, which uses keys and values from the first frame to generate the image of frame $k = 1, \ldots, m$. By using cross-frame attention, the appearance and the identity of the foreground object are preserved throughout the sequence. Optionally, we apply background smoothing. To this end, we employ salient object detection to obtain for each frame k a mask M^k indicating the foreground pixels. Finally, for the background (using the mask M^k), a convex combination between the latent code x_t^1 of frame one warped to frame k and the latent code x_t^k is used to further improve the temporal consistency of the background.

leading to temporal inconsistencies, especially for the foreground object.

3.3.2 Reprogramming Cross-Frame Attention

To address the issue mentioned above, we use a cross-frame attention mechanism to preserve the information about (in particular) the foreground object's appearance, shape, and identity throughout the generated video.

To leverage the power of cross-frame attention and at the same time exploit a pretrained SD without retraining, we replace each of its self-attention layers with a cross-frame attention, with the attention for each frame being on the first frame. More precisely in the original SD UNet architecture $\epsilon_\theta^t(x_t, \tau)$, each self-attention layer takes a feature map $x \in \mathbb{R}^{h \times w \times c}$, linearly projects it into query, key, value features $Q, K, V \in \mathbb{R}^{h \times w \times c}$, and computes the layer output by the following formula (for simplicity described here for only one attention head) [38]:

$$\text{Self-Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{c}}\right)V. \tag{7}$$

In our case, each attention layer receives m inputs: $x^{1:m} = [x^1, \ldots, x^m] \in \mathbb{R}^{m \times h \times w \times c}$. Hence, the linear projection layers produce m queries, keys, and values $Q^{1:m}, K^{1:m}, V^{1:m}$, respectively.

Therefore, we replace each self-attention layer with a cross-frame attention of each frame on the first frame as follows:

$$\begin{aligned}\text{Cross-Frame-Attn}(Q^k, K^{1:m}, V^{1:m}) = \\ \text{Softmax}\left(\frac{Q^k (K^1)^T}{\sqrt{c}}\right)V^1\end{aligned} \tag{8}$$

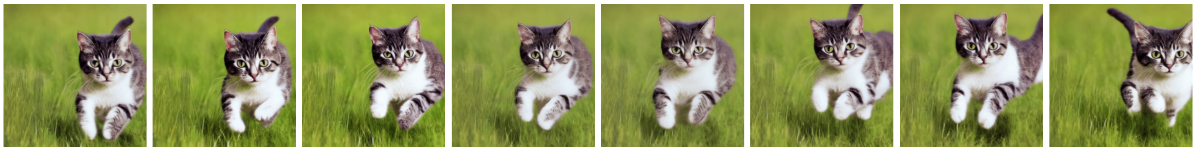
for $k = 1, \ldots, m$. By using cross frame attention, the appearance and structure of the objects and background as well as identities are carried over from the first frame to subsequent frames, which significantly increases the temporal consistency of the generated frames (see in the appendix the Figures 14, 16, 22, 23).

3.3.3 Background smoothing (Optional)

We further improve temporal consistency of the background using a convex combination of background-masked latent codes between the first frame and frame k. This helps especially to generate videos from textual prompts when one or no initial image and no further guidance are provided.

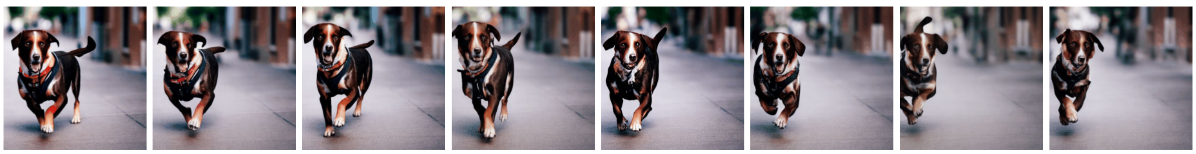
In more detail, given the generated sequence of our video generator, $x_0^{1:m}$, we apply (an in-house solution for) salient object detection [40] to the decoded images to obtain a corresponding foreground mask M^k for each frame k. Then we warp x_t^1 according to the employed motion dynamics defined by W_k and denote the result by $\hat{x}_t^k := W_k(x_t^1)$.

Background smoothing is achieved by a convex combination between the actual latent code x_t^k and the warped

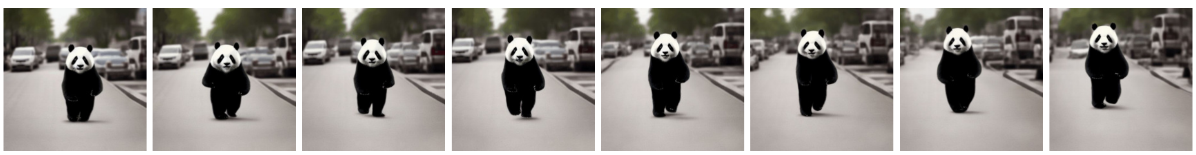


(a) a high quality realistic photo of a cute cat running in a beautiful meadow

(b) an astronaut is skiing down a hill



(c) a dog is walking down the street



(d) a high quality realistic photo of a panda walking alone down the street

Figure 3: Text-to-Video results of our method. Depicted frames show that identities and appearances are temporally consistent and fitting to the textual prompt. For more results, see Appendix Sec. 8.

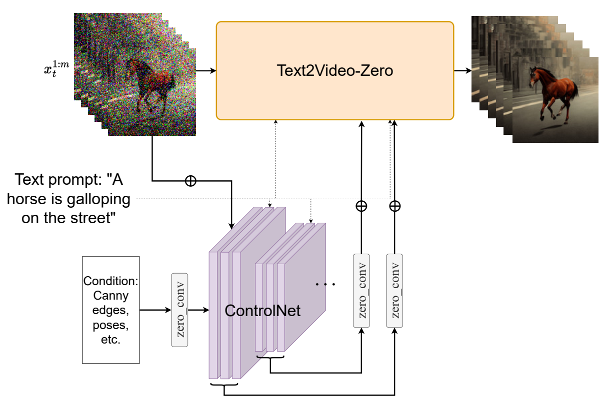


Figure 4: The overview of Text2Video-Zero + ControlNet

latent code $\hat{x}_t^k$ on the background, i.e.,

$$\overline{x}_t^k = M^k \odot x_t^k + (1 - M^k) \odot (\alpha \hat{x}_t^k + (1-\alpha) x_t^k), \quad (9)$$

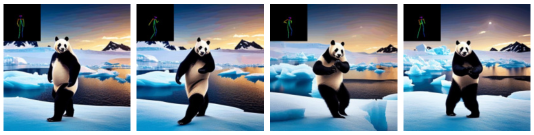
for $k = 1, \ldots, m$, where α is a hyperparameter, which we empirically choose $\alpha = 0.6$. Finally, DDIM sampling is employed on $\overline{x}_t^k$, which delivers video generation with background smoothing. We use background smoothing in our video generation from text when no guidance is provided. For an ablation study on background smoothing, see the appendix, Sec. 8.1.

3.4. Conditional and Specialized Text-to-Video

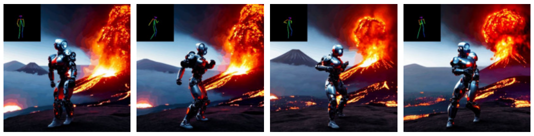
Recently powerful controlling mechanisms [47, 22, 17] emerged to guide the diffusion process for text-to-image generation. Particularly, ControlNet [47] enables to condition the generation process using edges, pose, semantic masks, image depths, etc. However, a direct application of ControlNet in the video domain leads to temporal inconsistencies and to severe changes of object appearance, identity, and the background (see in the appendix Figures 14, 16, 22, 23). It turns out that our modifications on the basic diffusion process for videos result in more consistent videos guided by ControlNet conditions. We would like to point out again that our method does not require any fine-tuning or optimization processes.

More specifically, ControlNet creates a trainable copy of the encoder (including the middle blocks) of the UNet $\epsilon_\theta^t(x_t, \tau)$ while additionally taking the input x_t and a condition c, and adds the outputs of each layer to the skip-connections of the original UNet. Here c can be any type of condition, such as edge map, scribbles, pose (body landmarks), depth map, segmentation map, etc. The trainable branch is being trained on a specific domain for each type of the condition c resulting in an effective conditional text-to-image generation mechanism.

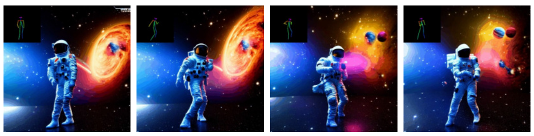
To guide our video generation process with ControlNet



(a) a panda dancing in Antarctica

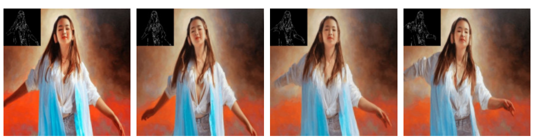


(b) a cyborg dancing near a volcano

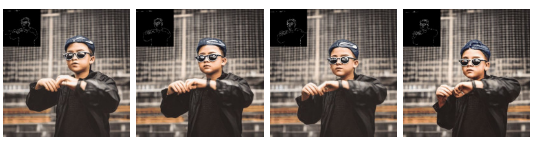


(c) an astronaut dancing in the outer space

Figure 5: Conditional generation with pose control. For more results see appendix, Sec. 10.



(a) oil painting of a girl dancing close-up



(b) Cyberpunk boy with a hat dancing close-up

Figure 6: Conditional generation with edge control. For more results see appendix, Sec. 9.

we apply our method to the basic diffusion process, i.e. enrich the latent codes $x_T^{1:m}$ with motion information and change the self-attentions into cross-frame attentions in the main UNet. While adopting the main UNet for video generation task, we apply the ControlNet pretrained copy branch per-frame on each x_t^k for $k = 1, \ldots, m$ in each diffusion time-step $t = T, \ldots, 1$ and add the ControlNet branch outputs to the skip-connections of the main UNet.

Furthermore, for our conditional generation task, we adopted the weights of specialized DreamBooth (DB) [31] models[1]. This gives us specialized time-consistent video generations (see Fig. 7).

[1] Avatar model: https://civitai.com/models/9968/avatar-style. GTA-5 model: https://civitai.com/models/1309/gta5-artwork-diffusion.

(a) oil painting of a beautiful girl avatar style

(b) gta-5 style

Figure 7: Conditional generation with edge control and DB models.

3.5. Video Instruct-Pix2Pix

With the rise of text-guided image editing methods such as Prompt2Prompt [9], Instruct-Pix2Pix [2], SDEdit [20], etc., text-guided video editing approaches emerged [1, 16, 42]. While these methods require complex optimization processes, our approach enables the adoption of any SD-based text-guided image editing algorithm to the video domain without any training or fine-tuning. Here we take the text-guided image editing method Instruct-Pix2Pix and combine it with our approach. More precisely, we change the self-attention mechanisms in Instruct-Pix2Pix to cross-frame attentions according to Eq. 8. Our experiments show that this adaptation significantly improves the consistency of the edited videos (see Fig. 9) over the naïve per-frame usage of Instruct-Pix2Pix.

4. Experiments

4.1. Implementation Details

We take the Stable Diffusion [29] code[2] with its pre-trained weights from version 1.5 as basis and implement our modifications.For each video, we generate $m = 8$ frames with 512×512 resolution. However, our framework allows generating any number of frames, either by increasing m, or by employing our method in an auto-regressive fashion where the last generated frame m becomes the first frame in computing the next m frames. For all text-to-video generation experiments, we take $T' = 881, T = 941$ without specific tuning, while for conditional and specialized generation, and for Video Instruct-Pix2Pix, we take $T' = T = 1000$.

For a conditional generation, we use the codebase[3] of ControlNet [47]. For specialized models, we take DB [31] models from publicly available sources. For Video Instruct-

[2] https://github.com/huggingface/diffusers. We also benefit from the codebase of Tune-A-Video https://github.com/showlab/Tune-A-Video.

[3] https://github.com/lllyasviel/ControlNet.

Pix2Pix, we use the codebase[4] of Instruct Pix2Pix [2].

4.2. Qualitative Results

All applications of Text2Video-Zero show that it successfully generates videos where the global scene and the background are time consistent and the context, appearance, and identity of the foreground object are maintained throughout the entire sequence.

In the case of text-to-video, we observe that it generates high-quality videos that are well-aligned to the text prompt (see Fig. 3 and the appendix). For instance, the depicted panda shows a naturally walking on the street. Likewise, using additional guidance from edges or poses (see Fig. 5, Fig, 6 and Fig. 7 and the appendix), high quality videos are generated matching the prompt and the guidance that show great temporal consistency and identity preservation.

Videos generated by Video Instruct-Pix2Pix (see Fig. 1 and the appendix) possess high-fidelity with respect to the input video, while following closely the instruction.

4.3. Comparison with Baselines

We compare our method with two publicly available baselines: CogVideo [15] and Tune-A-Video [42]. Since CogVideo is a text-to-video method we compare with it in pure text-guided video synthesis settings. With Tune-A-Video we compare in our Video Instruct-Pix2Pix setting.

4.3.1 Quantitative Comparison

To show quantitative results, we evaluate the CLIP score [10], which indicates video-text alignment. We randomly take 25 videos generated by CogVideo and synthesize corresponding videos using the same prompts according to our method. The CLIP scores for our method and CogVideo are 31.19 and 29.63, respectively. Our method thus slightly outperforms CogVideo, even though the latter has 9.4 billion parameters and requires large-scale training on videos.

4.3.2 Qualitative Comparison

We present several results of our method in Fig. 8 and provide a qualitative comparison to CogVideo [15]. Both methods show good temporal consistency throughout the sequence, preserving the identity of the object and background. However, our method shows better text-video alignment. For instance, while our method correctly generates a video of a man riding a bicycle in the sunshine in Fig. 8(b), CogVideo sets the background to moon light. Also in Fig. 8(a), our method correctly shows a man running in the snow, while neither the snow nor a man running are clearly visible in the video generated by CogVideo.

[4] https://github.com/timothybrooks/instruct-pix2pix.

Qualitative results of *Video Instruct-Pix2Pix* and a visual comparison with per-frame Instruct-Pix2Pix and Tune-A-Video are shown in Fig. 9. While Instruct-Pix2Pix shows a good editing performance per frame, it lacks temporal consistency. This becomes evident especially in the video depicting a skiing person, where the snow and the sky are drawn using different styles and colors. Using our Video Instruct-Pix2Pix method, these issues are solved resulting in temporally consistent video edits throughout the sequence.

While Tune-A-Video creates temporally consistent video generations, it is less aligned to the instruction guidance than our method, struggles creating local edits and losses details of the input sequence. This becomes apparent when looking at the edit of the dancer video depicted in Fig. 9 (left side). In contrast to Tune-A-Video, our method draws the entire dress brighter and at the same time better preserves the background, *e.g*. the wall behind the dancer is almost kept the same. Tune-A-Video draws a severely modified wall. Moreover, our method is more faithful to the input details, *e.g*., Video Instruct-Pix2Pix draws the dancer using the pose exactly as provided (Fig. 9 left), and shows all skiing persons appearing in the input video (compare last frame of Fig. 9(right)), in constrast to Tune-A-Video. All the above-mentioned weaknesses of Tune-A-Video can also be observed in our additional evaluations that are provided in the appendix, Figures 25, 26.

4.4. Ablation Study

We perform several ablation studies and provide the results in the appendix, Sect. 8.1. Namely, we analyse background smoothing, Δt, and two main components of our method: making the initial latent codes coherent to a motion, and using cross-frame attention on the first frame instead of self-attention.

5. Limitations and Future Work

The main limitation of this work is the inability to generate longer videos with sequences of actions. Future research may target enriching our method with techniques such as autoregressive scene action generation, while keeping the training-free spirit. Overall, our method is focused on generating video key-frames as in the first stage of Imagen Video [11] and Make-A-Video [33], and can thus be considered as good basis for longer and smoother video generation by integrating temporal upsampling.

6. Conclusion

In this paper, we addressed the problem of zero-shot text-to-video synthesis and proposed a novel method for time-consistent video generation. Our approach does not require any optimization or fine-tuning, making text-to-video generation and its applications affordable for everyone. We


(a) a man is running in the snow

(b) a man is riding a bicycle in the sunshine

(c) a man is walking in the rain

Figure 8: Comparison of our method vs CogVideo on text-to-video generation task (left is ours, right is CogVideo [15]). For more comparisons, see appendix Fig. 18.

Figure 9: Comparison of Video Instruct-Pix2Pix(ours) with Tune-A-Video and per-frame Instruct-Pix2Pix. For more comparisons see our appendix.

demonstrated the effectiveness of our method for various applications, including conditional and specialized video generation, and Video Instruct-Pix2Pix, *i.e.*, instruction-guided video editing. Our contributions to the field include presenting a new problem of zero-shot text-to-video synthesis, showing the utilization of text-to-image diffusion models for generating time-consistent videos, and providing evidence of the effectiveness of our method for various video synthesis applications. We believe that our proposed method will open up new possibilities for video generation and editing, making it accessible and affordable for everyone.

Acknowledgments. This material is partially based upon work supported by the National Science Foundation CAREER Award #2239840, and the National AI Institute for Exceptional Education (Award #2229873) by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022. 7

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 2, 7, 8, 13, 25, 26

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 4

[4] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 3

[5] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 2, 3, 4

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 3

[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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022. 2

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[9] 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. 7

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 8

[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. 2, 3, 4, 8

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 12

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2, 3, 4

[15] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3, 4, 8, 9, 12, 13, 20

[16] Yao-Chih Lee, Ji-Ze Genevieve Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing. *arXiv e-prints*, pages arXiv–2301, 2023. 7

[17] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 289–299, 2023. 6

[18] Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14267–14276, 2023. 3

[19] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015. 2

[20] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 7

[21] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 2

[22] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 6

[23] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

[24] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019. 2

[25] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 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. 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 *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[28] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. *Advances in neural information processing systems*, 29, 2016. 2

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 7

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[31] 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. 3, 7, 22

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[33] 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. *arXiv preprint arXiv:2209.14792*, 2022. 2, 3, 8

[34] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[37] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 5

[39] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 3, 4

[40] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3239–3259, 2021. 5

[41] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 720–736. Springer, 2022. 3, 4

[42] 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, 7, 8, 12, 13, 26

[43] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[44] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022. 2, 3

[45] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[46] 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

[47] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3, 6, 7