TI2V-Zero: Zero-Shot Image Conditioning for Text-to-Video Diffusion Models

Haomiao Ni[1*] Bernhard Egger[2] Suhas Lohit[3] Anoop Cherian[3] Ye Wang[3]
Toshiaki Koike-Akino[3] Sharon X. Huang[1] Tim K. Marks[3]
[1]The Pennsylvania State University, USA [2]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[3]Mitsubishi Electric Research Laboratories (MERL), USA
[1]{hfn5052, suh972}@psu.edu [2]bernhard.egger@fau.de [3]{slohit,cherian,yewang,koike,tmarks}@merl.com

https://merl.com/demos/TI2V-Zero

Abstract

Text-conditioned image-to-video generation (TI2V) aims to synthesize a realistic video starting from a given image (e.g., a woman's photo) and a text description (e.g., "a woman is drinking water."). Existing TI2V frameworks often require costly training on video-text datasets and specific model designs for text and image conditioning. In this paper, we propose TI2V-Zero, a zero-shot, tuning-free method that empowers a pretrained text-to-video (T2V) diffusion model to be conditioned on a provided image, enabling TI2V generation without any optimization, fine-tuning, or introducing external modules. Our approach leverages a pretrained T2V diffusion foundation model as the generative prior. To guide video generation with the additional image input, we propose a "repeat-and-slide" strategy that modulates the reverse denoising process, allowing the frozen diffusion model to synthesize a video frame-by-frame starting from the provided image. To ensure temporal continuity, we employ a DDPM inversion strategy to initialize Gaussian noise for each newly synthesized frame and a resampling technique to help preserve visual details. We conduct comprehensive experiments on both domain-specific and open-domain datasets, where TI2V-Zero consistently outperforms a recent open-domain TI2V model. Furthermore, we show that TI2V-Zero can seamlessly extend to other tasks such as video infilling and prediction when provided with more images. Its autoregressive design also supports long video generation.

1. Introduction

Image-to-video (I2V) generation is an appealing topic with various applications, including artistic creation, entertainment, and data augmentation for machine learning [39]. Given a single image x^0 and a text prompt y, text-conditioned image-to-video (TI2V) generation aims to synthesize M new frames to yield a realistic video, $\hat{\mathbf{x}} = \langle x^0, \hat{x}^1, \ldots, \hat{x}^M \rangle$, starting from the given frame x^0 and satisfying the text description y. Current TI2V generation methods [59, 63, 70] typically rely on computationally-heavy training on video-text datasets and specific architecture designs to enable text and image conditioning. Some [12, 25] are constrained to specific domains due to the lack of training with large-scale open-domain datasets. Other approaches, such as [14, 67], utilize pretrained foundation models to reduce training costs, but they still need to train additional modules using video data.

In this paper, we propose TI2V-Zero, which achieves *zero-shot* TI2V generation using only an open-domain pretrained text-to-video (T2V) latent diffusion model [60]. Here "zero-shot" means that when using the diffusion

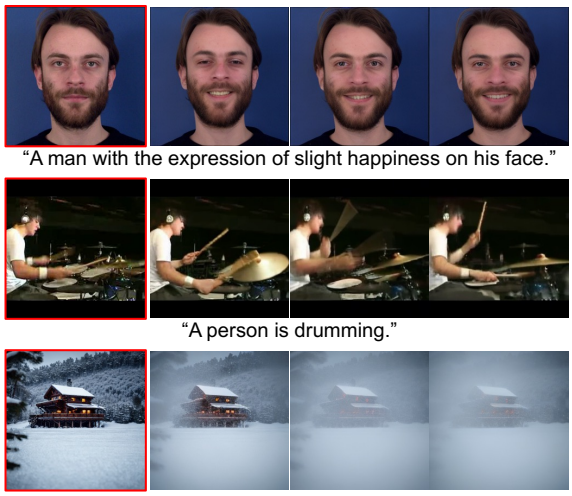

Figure 1. Examples of generated video frames using our proposed TI2V-Zero. The given first image x^0 is highlighted with the red box, and the text condition y is shown under each row of the video. The remaining columns show the 6th, 11th, and 16th frames of the generated output videos. Each generated video has 16 frames with a resolution of 256×256.

*Work done during an internship at MERL.

model (DM) that was trained only for text conditioning, our framework enables image conditioning without any optimization, fine-tuning, or introduction of additional modules. Specifically, we guide the generation process by incorporating the provided image x^0 into the output latent code at each reverse denoising step. To ensure that the temporal attention layers of the pretrained DM focus on information from the given image, we propose a "repeat-and-slide" strategy to synthesize the video in a frame-by-frame manner, rather than directly generating the entire video volume. Notably, TI2V-Zero is not trained for the specific domain of the provided image, thus allowing the model to generalize to any image during inference. Additionally, its autoregressive generation makes the synthesis of long videos possible.

While the standard denoising sampling process starting with randomly initialized Gaussian noise can produce matching semantics, it often results in temporally inconsistent videos. Therefore, we introduce an inversion strategy based on the DDPM [20] forward process, to provide a more suitable initial noise for generating each new frame. We also apply a resampling technique [33] in the video DM to help preserve the generated visual details. Our approach ensures that the network maintains temporal consistency, generating visually convincing videos conditioned on the given starting image (see Fig. 1).

We conduct extensive experiments on MUG [1], UCF-101 [56], and a new open-domain dataset. In these experiments, TI2V-Zero consistently performs well, outperforming a state-of-the-art model [67] that was based on a video diffusion foundation model [8] and was specifically trained to enable open-domain TI2V generation.

2. Related Work

2.1. Conditional Image-to-Video Generation

Conditional video generation aims to synthesize videos guided by user-provided signals. It can be classified according to which type(s) of conditions are given, such as text-to-video (T2V) generation [5, 16, 21, 23, 31, 65], video-to-video (V2V) generation [7, 38, 40, 45, 61, 64], and image-to-video (I2V) generation [4, 10, 25, 34, 39, 69]. Here we discuss previous text-conditioned image-to-video (TI2V) generation methods [12, 14, 22, 44, 63, 70]. Hu *et al*. [25] introduced MAGE, a TI2V generator that integrates a motion anchor structure to store appearance-motion-aligned representations through three-dimensional axial transformers. Yin *et al*. [70] proposed DragNUWA, a diffusion-based model capable of generating videos controlled by text, image, and trajectory information with three modules including a trajectory sampler, a multi-scale fusion, and an adaptive training strategy. However, these TI2V frameworks require computationally expensive training on video-text datasets and a particular model design to support text-and-image-conditioned training. In contrast, our proposed TI2V-Zero leverages a pretrained T2V diffusion model to achieve zero-shot TI2V generation without additional optimization or fine-tuning, making it suitable for a wide range of applications.

2.2. Adaptation of Diffusion Foundation Models

Due to the recent successful application of diffusion models (DM) [20, 42, 47, 54, 55] to both image and video generation, visual diffusion foundation models have gained prominence. These include text-to-image (T2I) models such as Imagen [50] and Stable Diffusion [47], as well as text-to-video (T2V) models such as ModelScopeT2V [60] and VideoCrafter1 [8]. These models are trained with large-scale open-domain datasets, often including LAION-400M [52] and WebVid-10M [2]. They have shown immense potential for adapting their acquired knowledge base to address a wide range of downstream tasks, thereby reducing or eliminating the need for extensive labeled data. For example, previous works have explored the application of large T2I models to personalized image generation [13, 49], image editing [17, 33, 35–37], image segmentation [3, 68], video editing [45, 62], and video generation [14, 27, 53, 66]. In contrast to T2I models, there are fewer works on the adaptation of large-scale T2V models. Xing *et al*. [67] proposed DynamicCrafter for open-domain TI2V generation by adapting a T2V foundation model [8]. To control the generative process, they first employed a learnable image encoding network to project the given image into a text-aligned image embedding space. Subsequently, they utilized dual cross-attention layers to fuse text and image information and also concatenated the image with the initial noise to provide the video DM with more precise image details. In contrast, in this paper we explore how to inject the provided image to guide the DM sampling process based solely on the pretrained T2V model itself, with no additional training for the new TI2V task.

3. Methodology

Given one starting image x^0 and text y, let $\mathbf{x} = \langle x^0, x^1, \ldots, x^M \rangle$ represent a real video corresponding to text y. The objective of text-conditioned image-to-video (TI2V) generation is to synthesize a video $\hat{\mathbf{x}} = \langle x^0, \hat{x}^1, \ldots, \hat{x}^M \rangle$, such that the conditional distribution of $\hat{\mathbf{x}}$ given x^0 and y is identical to the conditional distribution of $\mathbf{x}$ given x^0 and y, i.e., $p(\hat{\mathbf{x}}|x^0, y) = p(\mathbf{x}|x^0, y)$. Our proposed TI2V-Zero can be built on a pretrained T2V diffusion model with a 3D-UNet-based denoising network. Here we choose ModelScopeT2V [60] as backbone due to its promising open-domain T2V generation ability. Below, we first introduce preliminaries about diffusion models, then introduce the architecture of the pretrained T2V model, and finally present the details of our TI2V-Zero.

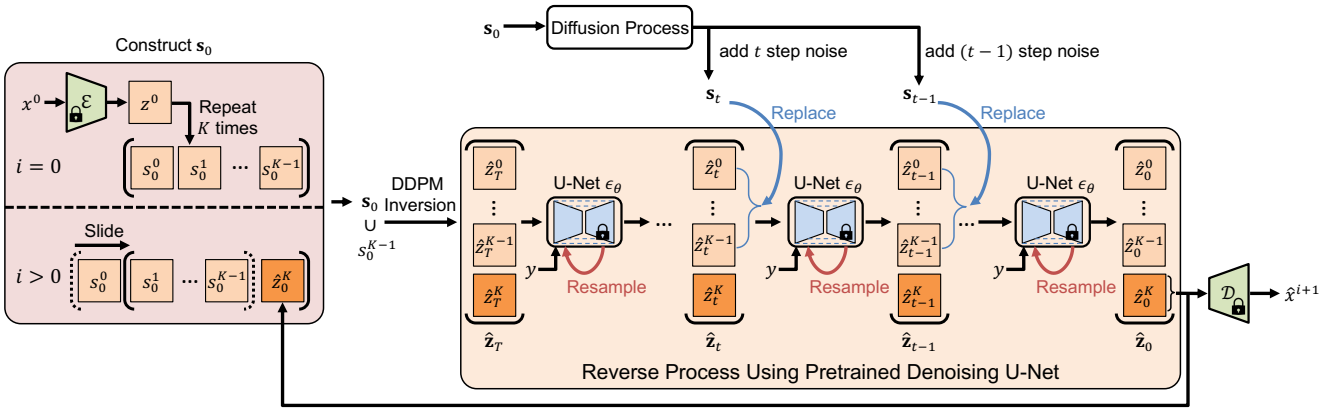

Figure 2. Illustration of the process of applying TI2V-Zero to generate the new frame $\hat{x}^{i+1}$, given the starting image x^0 and text y. TI2V-Zero is built upon a *frozen* pretrained T2V diffusion model, including frame encoder $\mathcal{E}$, frame decoder $\mathcal{D}$, and the denoising U-Net ϵ_θ. At the beginning of generation ($i = 0$), we encode x^0 as z^0 and repeat it K times to form the queue $\mathbf{s}_0$. We then apply DDPM-based inversion to $\mathbf{s}_0$ to produce the initial Gaussian noise $\hat{\mathbf{z}}_T$. Subsequently, in each reverse denoising step using U-Net ϵ_θ, we keep replacing the first K frames of $\hat{\mathbf{z}}_t$ with the noisy latent code $\mathbf{s}_t$ derived from $\mathbf{s}_0$. Resampling is also applied within each step to improve motion coherence. We finally decode the final frame of the clean latent code $\hat{\mathbf{z}}_0$ as the new synthesized frame $\hat{x}^{i+1}$. To compute the new $\mathbf{s}_0$ for the next iteration of generation ($i > 0$), we perform a sliding operation by dequeuing s_0^0 and enqueuing $\hat{z}_0^K$ within $\mathbf{s}_0$.

3.1. Preliminaries: Diffusion Models

Diffusion Models (DM) [20, 54, 55] are probabilistic models designed to learn a data distribution. Here we introduce the fundamental concepts of Denoising Diffusion Probabilistic Models (DDPM). Given a sample from the data distribution $\mathbf{z}_0 \sim q(\mathbf{z}_0)$, the *forward* diffusion process of a DM produces a Markov chain $\mathbf{z}_1, \ldots, \mathbf{z}_T$ by iteratively adding Gaussian noise to $\mathbf{z}_0$ according to a variance schedule $\beta_1, \ldots, \beta_T$, that is:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}) \ , \quad (1)$$

where variances β_t are constant. When the β_t are small, the posterior $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$ can be well approximated by a diagonal Gaussian [41, 54]. Furthermore, if the length of the chain, denoted by T, is sufficiently large, $\mathbf{z}_T$ can be well approximated by a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. These suggest that the true posterior $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$ can be estimated by $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$ defined as:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t), \sigma_t^2\mathbf{I}) \ , \quad (2)$$

where variances σ_t are also constants. The *reverse* denoising process in the DM (also termed *sampling*) then generates samples $\mathbf{z}_0 \sim p_\theta(\mathbf{z}_0)$ by starting with Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and gradually reducing noise in a Markov chain $\mathbf{z}_{T-1}, \mathbf{z}_{T-2}, \ldots, \mathbf{z}_0$ using a learned $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$. To learn $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$, Gaussian noise ϵ is first added to $\mathbf{z}_0$ to generate samples $\mathbf{z}_t$. Utilizing the independence property of the noise added at each forward step in Eq. (1), we can calculate the total noise variance as $\bar{\alpha}_t = \prod_{i=0}^{t}(1-\beta_i)$ and transform $\mathbf{z}_0$ to $\mathbf{z}_t$ in a single step:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}) \ . \quad (3)$$

Then a model ϵ_θ is trained to predict ϵ using the following mean-squared error loss:

$$L = \mathbb{E}_{t\sim\mathcal{U}(1,T),\mathbf{z}_0\sim q(\mathbf{z}_0),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t)||^2\right] \ , \quad (4)$$

where diffusion step t is uniformly sampled from $\{1, \ldots, T\}$. Then $\mu_\theta(\mathbf{z}_t)$ in Eq. (2) can be derived from $\epsilon_\theta(\mathbf{z}_t, t)$ to model $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$ [20]. The denoising model ϵ_θ is implemented using a time-conditioned U-Net [48] with residual blocks [15] and self-attention layers [58]. Diffusion step t is specified to ϵ_θ by the sinusoidal position embedding [58]. Conditional generation that samples $\mathbf{z}_0 \sim p_\theta(\mathbf{z}_0|y)$ can be achieved by learning a y-conditioned model $\epsilon_\theta(\mathbf{z}_t, t, y)$ [41, 47] with *classifier-free* guidance [19]. During training, the condition y in $\epsilon_\theta(\mathbf{z}_t, t, y)$ is replaced by a null label $\emptyset$ with a fixed probability. When sampling, the output is generated as follows:

$$\hat{\epsilon}_\theta(\mathbf{z}_t, t, y) = \epsilon_\theta(\mathbf{z}_t, t, \emptyset) + g \cdot (\epsilon_\theta(\mathbf{z}_t, t, y) - \epsilon_\theta(\mathbf{z}_t, t, \emptyset)) \ , \quad (5)$$

where g is the guidance scale.

3.2. Architecture of Pretrained T2V Model

TI2V-Zero can be built upon a pretrained T2V diffusion model with a 3D-UNet-based denoising network. Here we choose ModelScopeT2V [60] as the pretrained model (denoted $\mathcal{M}$). We now describe this T2V model in detail.

Structure Overview. Given a text prompt y, the T2V model $\mathcal{M}$ synthesizes a video $\hat{\mathbf{x}} = \langle \hat{x}^0, \hat{x}^1, \ldots, \hat{x}^K \rangle$ with a pre-defined video of length $(K+1)$ using a latent video diffusion model. Similar to Latent Diffusion Models (LDM) [47], $\mathcal{M}$ incorporates a frame auto-encoder [11, 28] for the conversion of data between pixel space $\mathcal{X}$ and latent space $\mathcal{Z}$ through its encoder $\mathcal{E}$ and decoder $\mathcal{D}$. Given the real video $\mathbf{x} = \langle x^0, x^1, \ldots, x^K \rangle$, $\mathcal{M}$ first utilizes the frame encoder $\mathcal{E}$ to encode the video $\mathbf{x}$ as $\mathbf{z} = \langle z^0, z^1, \ldots, z^K \rangle$. Here the sizes of pixel frame x and latent frame z are $H_x \times W_x \times 3$ and $H_z \times W_z \times C_z$, respectively. To be consistent with the notation used for the DM, we denote the

Algorithm 1 Generation using our TI2V-Zero approach.

Input: The starting frame x^0; The text prompt y; The pretrained T2V Model $\mathcal{M}$ for generating $(K+1)$-frame videos, including frame encoder $\mathcal{E}$ and frame decoder $\mathcal{D}$, and the DM denoising networks ϵ_θ; The iteration number U for resampling; The parameter M to control the length of the output video.
Output: A synthesized video $\hat{\mathbf{x}}$ with $(M+1)$ frames.

$z^0 \leftarrow \mathcal{E}(x^0)$ // Encode x^0
$\mathbf{s}_0 \leftarrow \langle z^0, z^0, \cdots, z^0 \rangle$ // Repeat z^0 for K times
$\hat{\mathbf{x}} \leftarrow \langle x^0 \rangle$
for $i = 1, 2, \cdots, M$ **do**
  // Generate one new frame $\hat{x}^i$
  $\mathbf{s}_T \sim \mathcal{N}(\sqrt{\bar{\alpha}_T}\mathbf{s}_0, (1-\bar{\alpha}_T)\mathbf{I})$ // DDPM Inversion
  $\hat{z}_T^K \sim \mathcal{N}(\sqrt{\bar{\alpha}_T}s_0^{K-1}, (1-\bar{\alpha}_T)\mathbf{I})$
  $\hat{\mathbf{z}}_T \leftarrow \mathbf{s}_T \cup \hat{z}_T^K$ // Initialize $\hat{\mathbf{z}}_T$
  for $t = T-1, \cdots, 2, 1$ **do**
    $\mathbf{s}_t \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{s}_0, (1-\bar{\alpha}_t)\mathbf{I})$
    for $u = 1, 2, \cdots, U$ **do**
      $\langle \hat{z}_t^0, \hat{z}_t^1, \cdots, \hat{z}_t^{K-1} \rangle \leftarrow \mathbf{s}_t$ // Replace
      $\hat{\mathbf{z}}_{t-1} \sim \mathcal{N}(\mu_\theta(\hat{\mathbf{z}}_t, y), \sigma_t^2\mathbf{I})$
      if $u < U$ and $t > 1$ **then**
        $\hat{\mathbf{z}}_t \sim \mathcal{N}(\sqrt{1-\beta_t}\hat{\mathbf{z}}_{t-1}, \beta_t\mathbf{I})$ // Resample
      end if
    end for
  end for
  $\mathbf{s}_0 \leftarrow \langle s_0^1, s_0^2, \cdots, s_0^{K-1} \rangle \cup \hat{z}_0^K$ // Slide
  $\hat{x}^i \leftarrow \mathcal{D}(\hat{z}_0^K)$ // Decode $\hat{z}_0^K$
  $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} \cup \hat{x}^i$
end for
return $\hat{\mathbf{x}}$

clean video latent $\mathbf{z} = \mathbf{z}_0 = \langle z_0^0, z_0^1, \ldots, z_0^K \rangle$. $\mathcal{M}$ then learns a DM on the latent space $\mathcal{Z}$ through a 3D denoising U-Net ϵ_θ [9]. Let $\mathbf{z}_t = \langle z_t^0, z_t^1, \ldots, z_t^K \rangle$ represent the latent sequence that results from adding noise over t steps to the original latent sequence $\mathbf{z}_0$. When training, the forward diffusion process of a DM transforms the initial latent sequence $\mathbf{z}_0$ into $\mathbf{z}_T$ by iteratively adding Gaussian noise ϵ for T steps. During inference, denoising U-Net ϵ_θ predicts the added noise at each step, enabling the generation of the clean latent sequence $\hat{\mathbf{z}}_0 = \langle \hat{z}_0^0, \hat{z}_0^1, \ldots, \hat{z}_0^K \rangle$ starting from randomly sampled Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Text Conditioning Mechanism. $\mathcal{M}$ employs a cross-attention mechanism [47] to incorporate text information into the generative process as guidance. Specifically, $\mathcal{M}$ uses a pretrained CLIP model [46] to encode the prompt y as the text embedding e. The embedding e is later used as the key and value in the multi-head attention layer within the spatial attention blocks, thus enabling the integration of text features with the intermediate U-Net features in ϵ_θ.

Denoising U-Net. The denoising U-Net ϵ_θ includes four key building blocks: the initial block, the downsampling block, the spatio-temporal block, and the upsampling block. The initial block transfers the input into the embedding space, while the downsampling and upsampling blocks are responsible for spatially downsampling and upsampling the feature maps. The spatio-temporal block is designed to capture spatial and temporal dependencies in the latent space, which comprises 2D spatial convolution, 1D temporal convolution, 2D spatial attention, and 1D temporal attention.

Figure 3. Illustration of the motivation behind our framework. We explore the application of a replacing-based baseline approach (rows 2–4, labeled "Replacing") and our TI2V-Zero (rows 5–6, labeled "TI2V-Zero") in various video generation tasks. The given real frames for each task are highlighted by red boxes and the text input is shown under the block. The replacing-based approach is only effective at predicting a single frame when all the other frames in the video are provided, while TI2V-Zero generates temporally coherent videos for both the TI2V and video infilling tasks.

3.3. Our Framework

Leveraging the pretrained T2V foundation model $\mathcal{M}$, we first propose a straightforward *replacing*-based baseline for adapting $\mathcal{M}$ to TI2V generation. We then analyze the possible reasons why it fails and introduce our TI2V-Zero framework, which includes a repeat-and-slide strategy, DDPM-based inversion, and resampling. Figure 2 and Algorithm 1 demonstrate the inference process of TI2V-Zero.

Replacing-based Baseline. We assume that the pretrained model $\mathcal{M}$ is designed to generate the video with a fixed length of $(K+1)$. So we first consider synthesizing videos with that same length $(K+1)$, i.e., $M = K$. Since the DM process operates within the latent space $\mathcal{Z}$, we use the encoder $\mathcal{E}$ to map the given starting frame x^0 into the latent representation z^0. Additionally, we denote $z^0 = z_0^0$ to specify that the latent is clean and corresponds

to diffusion step 0 of the DM. Note that each reverse denoising step in Eq. (2) from $\hat{\mathbf{z}}_t$ to $\hat{\mathbf{z}}_{t-1}$ depends solely on $\hat{\mathbf{z}}_t = \langle \hat{z}_t^0, \hat{z}_t^1, \ldots, \hat{z}_t^K \rangle$. To ensure that the first frame of the final synthesized clean video latent $\hat{\mathbf{z}}_0 = \langle \hat{z}_0^0, \hat{z}_0^1, \ldots, \hat{z}_0^K \rangle$ at step 0 matches the provided image latent, i.e., $\hat{z}_0^0 = z_0^0$, we can modify the first generated latent $\hat{z}_t^0$ of $\hat{\mathbf{z}}_t$ at each reverse step, as long as the signal-to-noise ratio of each frame latent in $\hat{\mathbf{z}}_t$ remains consistent. Using Eq. (3), we can add t steps of noise to the provided image latent z_0^0, allowing us to sample z_t^0 through a single-step calculation. By *replacing* the first generated latent $\hat{z}_t^0$ with the noisy image latent z_t^0 at each reverse denoising step, we might expect that the video generation process can be guided by z_0^0 with the following expressions defined for each reverse step:

$$z_t^0 \sim \mathcal{N}(\sqrt{\bar{\alpha}_t} z_0^0, (1 - \bar{\alpha}_t)\mathbf{I}) \ , \tag{6a}$$

$$\hat{z}_t^0 \leftarrow z_t^0 \ , \tag{6b}$$

$$\hat{\mathbf{z}}_{t-1} \sim \mathcal{N}(\mu_\theta(\hat{\mathbf{z}}_t, y), \sigma_t^2 \mathbf{I}) \ . \tag{6c}$$

Specifically, in each reverse step from $\hat{\mathbf{z}}_t$ to $\hat{\mathbf{z}}_{t-1}$, as shown in Eq. (6a), we first compute the noisy latent z_t^0 by adding Gaussian noise to the given image latent z_0^0 over t steps. Then, we replace the first latent $\hat{z}_t^0$ of $\hat{\mathbf{z}}_t$ with z_t^0 in Eq. (6b) to incorporate the provided image into the generation process. Finally, in Eq. (6c), we pass $\hat{\mathbf{z}}_t$ through the denoising network to generate $\hat{\mathbf{z}}_{t-1}$, where the text y is integrated by classifier-free guidance (Eq. (5)). After T iterations, the final clean latent $\hat{\mathbf{z}}_0$ at diffusion step 0 can be mapped back into the image space $\mathcal{X}$ using the decoder $\mathcal{D}$.

Using this replacing-based baseline, we might expect that the temporal attention layers in ϵ_θ can utilize the context provided by the first frame latent $\hat{z}_t^0$ to generate the subsequent frame latents in a manner that harmonizes with $\hat{z}_t^0$. However, as shown in Fig. 3, row 2, this replacing-based approach fails to produce a video that is temporally consistent with the first image. The generated frames are consistent with each other, but not with the provided first frame.

To analyze possible reasons for failure, we apply this baseline to a simpler video infilling task, where every other frame is provided and the model needs to predict the interspersed frames. In this case, the baseline replaces the generated frame latents at positions corresponding to real frames with noisy provided-frame latents in each reverse step. The resulting video, in Fig. 3, row 3, looks like a combination of two independent videos: the generated (even) frames are consistent with each other but not with the provided (odd) frames. We speculate that this may result from the intrinsic dissimilarity between frame latents derived from the given real images and those sampled from ϵ_θ. Thus, the temporal attention values between frame latents sampled in the same way (both from the given images or both from ϵ_θ) will be higher, while the attention values between frame latents sampled in different ways (one from the given image and the other from ϵ_θ) will be lower. Therefore, the temporal attention layers of $\mathcal{M}$ tend to utilize the information from latents

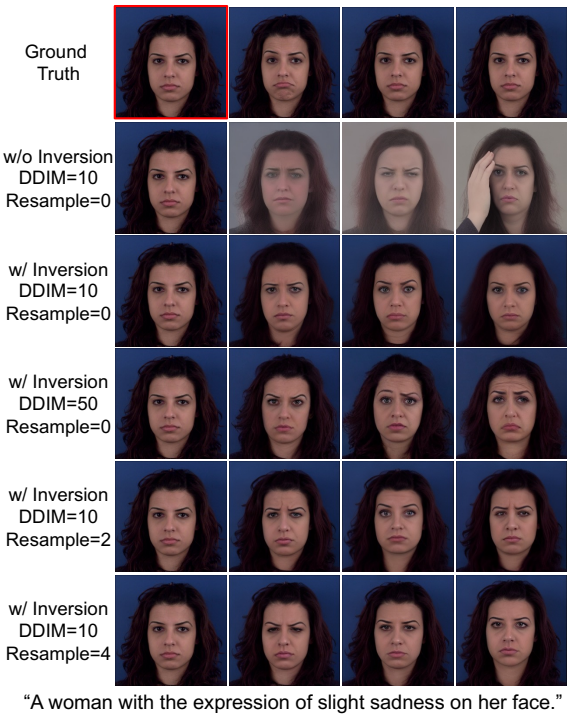

Figure 4. Qualitative ablation study comparing different sampling strategies for our TI2V-Zero on MUG. The first image $\hat{x}^0$ is highlighted with the red box and text y is shown under the block. The 1st, 6th, 11th, and 16th frames of the videos are shown in each column. The terms `Inversion`, `DDIM`, and `Resample` denote the application of DDPM inversion, the steps using DDIM sampling, and the iteration number using resampling, respectively.

produced by ϵ_θ to synthesize new frames at each reverse step, ignoring the provided frames. We further simplify the task to single-frame prediction, where the model only needs to predict a single frame when all the other frames in the video are given. In this setting, all the frame latents except for the final frame are replaced by noisy provided-frame latents in each reverse step. Thus, temporal attention layers can only use information from the real frames. In this case, Fig. 3, row 4, shows that the baseline can now generate a final frame that is consistent with the previous frames.

Repeat-and-Slide Strategy. Inspired by the observation in Fig. 3, to guarantee that the temporal attention layers of $\mathcal{M}$ depend solely on the given image, we make two major changes to the proposed replacing-based baseline: (1) instead of using $\mathcal{M}$ to directly synthesize the entire $(K+1)$-frame video, we switch to a frame-by-frame generation approach, i.e., we generate only one new frame latent in each complete DM sampling process; (2) for each sampling process generating the new frame latent, we ensure that only one frame latent is produced from ϵ_θ, while the other K frame latents are derived from the given real image and previously synthesized frames, thereby forcing temporal attention layers to only use the information from these frame latents. Specifically, we construct a queue of

K frame latents, denoted as $\mathbf{s}_0 = \langle s_0^0, s_0^1, \cdots, s_0^{K-1} \rangle$. We also define $\mathbf{s}_t = \langle s_t^0, s_t^1, \cdots, s_t^{K-1} \rangle$, which is obtained by adding t steps of Gaussian noise to the clean $\mathbf{s}_0$. Similar to our replacing-based baseline in the single-frame prediction task, in each reverse step from $\hat{\mathbf{z}}_t$ to $\hat{\mathbf{z}}_{t-1}$, we replace the first K frame latents in $\hat{\mathbf{z}}_t$ by $\mathbf{s}_t$. Consequently, the temporal attention layers have to utilize information from $\mathbf{s}_0$ to synthesize the new frame's latent, $\hat{z}_0^K$. Considering that only one starting image latent z^0 is provided, we propose a "repeat-and-slide" strategy to construct $\mathbf{s}_0$. At the beginning of video generation, we *repeat* z^0 for K frames to form $\mathbf{s}_0$, and gradually perform a *sliding* operation within the queue $\mathbf{s}_0$ by dequeuing the first frame latent s_0^0 and enqueuing the newly generated latent $\hat{z}_0^K$ after each complete DM sampling process. Note that though the initial $\mathbf{s}_0$ is created by repeating z^0, the noise added to get $\mathbf{s}_t$ is different for each frame's latent in $\mathbf{s}_t$, thus ensuring diversity. The following expressions define one reverse step in the DM sampling process:

$$\mathbf{s}_t \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{s}_0, (1-\bar{\alpha}_t)\mathbf{I}) \ , \tag{7a}$$

$$\langle \hat{z}_t^0, \hat{z}_t^1, \cdots, \hat{z}_t^{K-1} \rangle \leftarrow \mathbf{s}_t \ , \tag{7b}$$

$$\hat{\mathbf{z}}_{t-1} \sim \mathcal{N}(\mu_\theta(\hat{\mathbf{z}}_t, y), \sigma_t^2\mathbf{I}) \ . \tag{7c}$$

Specifically, in each reverse denoising step from $\hat{\mathbf{z}}_t$ to $\hat{\mathbf{z}}_{t-1}$, we first add t steps of Gaussian noise to the queue $\mathbf{s}_0$ to yield $\mathbf{s}_t$ in Eq. (7a). Subsequently, we replace the previous K frames of $\hat{\mathbf{z}}_t$ with $\mathbf{s}_t$ in Eq. (7b) and input $\hat{\mathbf{z}}_t$ to the denoising network to produce the less noisy latent $\hat{\mathbf{z}}_{t-1}$ (Eq. (7c)).

With the repeat-and-slide strategy, model $\mathcal{M}$ is tasked with predicting only one new frame, while the preceding K frames are incorporated into the reverse process to ensure that the temporal attention layers depend solely on information derived from the provided image.

DDPM-based Inversion. Though the DM sampling process starting with randomly sampled Gaussian noise produces matching semantics, the generated video is often temporally inconsistent (Fig. 4, row 2). To provide initial noise that can produce more temporally consistent results, we introduce an inversion strategy based on the DDPM [20] forward process when generating the new frame latent. Specifically, at the beginning of each DM sampling process to synthesize the new frame latent $\hat{z}_0^K$, instead of starting with the $\hat{\mathbf{z}}_T$ randomly sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, we add T full steps of Gaussian noise to $\mathbf{s}_0$ to obtain $\mathbf{s}_T$ using Eq. (3). Note that $\hat{\mathbf{z}}$ has $K+1$ frames, while $\mathbf{s}$ has K frames. We then use $\mathbf{s}_T$ to initialize the first K frames of $\hat{\mathbf{z}}_T$. We copy the last frame s_T^{K-1} of $\mathbf{s}_T$ to initialize the final frame $\hat{z}_T^K$, as the $(K-1)$th frame is the closest to the Kth frame.

Resampling. Similar to [24, 33], we further apply a resampling technique, which was initially designed for the image inpainting task, to the video DM to enhance motion coherence. Particularly, after performing a one-step denoising operation in the reversed process, we add one-step noise again to revert the latent. This procedure is repeated multiple times for each diffusion step, ensuring harmonization between the predicted and conditioning frame latents (see Algorithm 1 for details).

Inversion	DDIM	Resample	FVD↓	sFVD↓	tFVD↓
✗	10	0	1656.37	2074.77±411.74	1798.05±235.34
✓	10	0	339.89	443.97±139.10	405.22±61.58
✓	50	0	463.55	581.32±234.09	535.06±85.27
✓	10	2	207.62	299.14±87.24	278.73±47.84
✓	10	4	**180.09**	**267.17±74.72**	**252.77±39.02**

Table 1. Quantitative ablation study comparing different sampling strategies for proposed TI2V-Zero on the MUG dataset. `Inversion`, `DDIM`, and `Resample` denote the application of DDPM-based inversion, the steps using DDIM sampling, and the iteration number using resampling, respectively.

Distributions for Comparison	FVD↓	tFVD↓
TI2V-Zero-Fake *vs.* ModelScopeT2V	**366.41**	**921.31±251.85**
TI2V-Zero-Real *vs.* Real Videos	477.19	1306.75±271.82
ModelScopeT2V *vs.* Real Videos	985.82	2264.08±501.28
TI2V-Zero-Fake *vs.* Real Videos	**937.11**	**2177.70±436.71**

Table 2. Result analysis of TI2V-Zero starting from the real (i.e., TI2V-Zero-Real) or synthesized frames (i.e., TI2V-Zero-Fake) on the UCF101 dataset.

4. Experiments

4.1. Datasets and Metrics

We conduct comprehensive experiments on three datasets. More details about datasets, such as selected subjects and text prompts, can be found in our Supplementary Materials.

MUG facial expression dataset [1] contains 1,009 videos of 52 subjects performing 7 different expressions. We include this dataset to evaluate the performance of models in scenarios with small motion and a simple, unchanged background. To simplify the experiments, we randomly select 5 male and 5 female subjects, and 4 expressions. We use the text prompt templates like "`A woman with the expression of slight {label} on her face.`" to change the expression class label to be text input. Since the expressions shown in the videos of MUG are often not obvious, we add "slight" in the text input to avoid large motion.

UCF101 action recognition dataset [56] contains 13,320 videos from 101 human action classes. We include this dataset to measure performance under complicated motion and complex, changing backgrounds. To simplify the experiments, we select 10 action classes and the first 10 subjects within each class. We use text prompt templates such as "`A person is performing {label}.`" to change the class label to text input.

In addition to the above two datasets, we create an **OPEN** dataset to assess the model's performance in open-domain TI2V generation. We first utilize ChatGPT [43] to generate 10 text prompts. Subsequently, we employ Stable

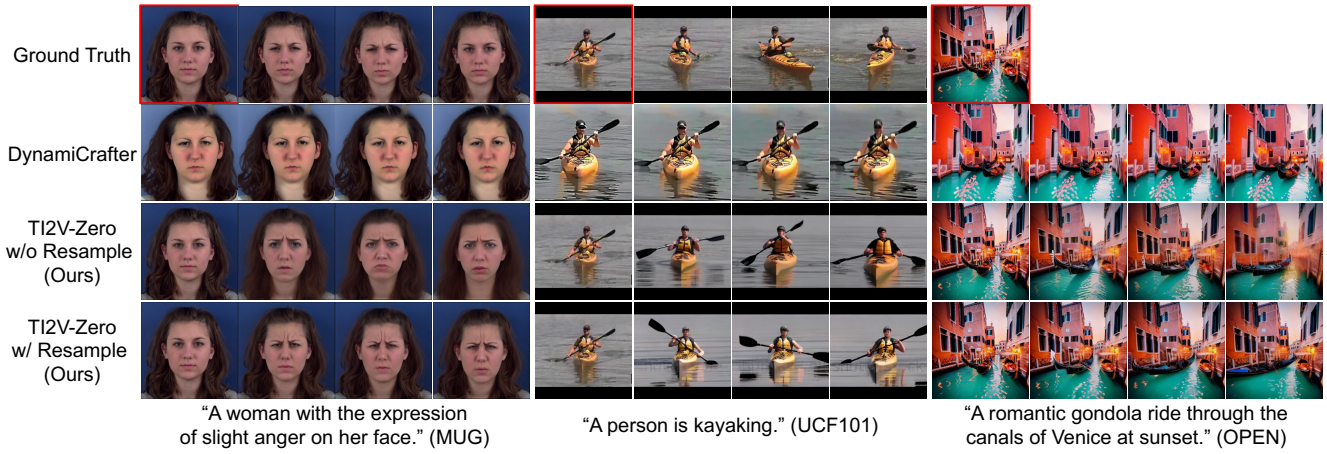

Figure 5. Qualitative comparison among different methods on multiple datasets for TI2V generation. Columns in each block display the 1st, 6th, 11th, and 16th frames of the output videos, respectively. There are 16 frames with a resolution of 256×256 for each video. The given image x^0 is highlighted with the red box and the text prompt y is shown under each block.

Model	MUG			UCF101	
	FVD↓	sFVD↓	tFVD↓	FVD↓	tFVD↓
DynamiCrafter [67]	1094.72	1359.86±257.73	1223.89±105.94	589.59	1540.02±199.59
TI2V-Zero w/o Resample (Ours)	339.89	443.97±139.10	405.22±61.58	493.19	1319.77±283.87
TI2V-Zero w/ Resample (Ours)	**180.09**	**267.17±74.72**	**252.77±39.02**	**477.19**	**1306.75±271.82**

Table 3. Quantitative comparison among different methods on multiple datasets for TI2V generation.

Diffusion 1.5 [47] to synthesize 100 images from each text prompt, generating a total of 1,000 starting images and 10 text prompts for evaluating TI2V models.

Data Preprocessing. We resize all the videos/images to 256×256 resolution. For UCF101, since most of the video frames are not square, we crop the central part of the frames. To obtain ground truth videos for computing metrics, we uniformly sample 16 frames from each video in the datasets to generate the video clips with a fixed length.

Metrics. Following prior work [21, 22, 25], we assess the *visual quality*, *temporal coherence*, and *sample diversity* of generated videos using Fréchet Video Distance (**FVD**) [57]. Similar to Fréchet Inception Distance (FID) [18], which is used for image quality evaluation, FVD utilizes a video classification network I3D [6] pretrained on Kinetics-400 dataset [26] to extract feature representation of real and synthesized videos. Then it calculates the Fréchet distance between the distributions of the real and synthesized video features. To measure how well a generated video aligns with the text prompt y (*condition accuracy*) and the given image x_0 (*subject relevance*), following [39], we design two variants of FVD, namely text-conditioned FVD (**tFVD**) and subject-conditioned FVD (**sFVD**). tFVD and sFVD compare the distance between real and synthesized video feature distributions under the same text y or the same subject image x_0, respectively. We first compute tFVD and sFVD for each condition y and image x_0, then report their mean and variance as final results. In our experiments, we generate 1,000 videos for all the models to estimate the feature distributions. We compute both tFVD and sFVD on the MUG dataset, but for UCF101, we only consider tFVD since it doesn't contain videos of different actions for the same subject. For the OPEN dataset, we only present qualitative results due to the lack of ground truth videos. Unless otherwise specified, all the generated videos are 16 frames (i.e., $M = 15$) with resolution 256×256.

4.2. Implementation Details

Model Implementation. We take the ModelScopeT2V 1.4.2 [60] as basis and implement our modifications. For text-conditioned generation, we employ classifier-free guidance with $g = 9.0$ in Eq. (5). Determined by our preliminary experiments, we choose 10-step DDIM and 4-step resampling as the default setting for MUG and OPEN, and 50-step DDIM and 2-step resampling for UCF101.

Implementation of SOTA Model. We compare our TI2V-Zero with a state-of-the-art (SOTA) model *DynamiCrafter*, a recent open-domain TI2V framework [67]. DynamiCrafter is based on a large-scale pretrained T2V foundation model VideoCrafter1 [16]. It introduces a learnable projection network to enable image-conditioned generation and then fine-tunes the entire framework. We implement DynamiCrafter using their provided code with their default settings. For a fair comparison, all the generated videos are

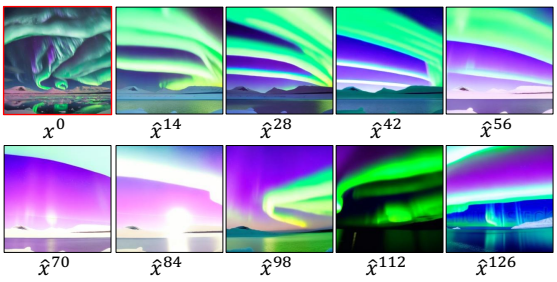

Figure 6. Example of long video generation using our TI2V-Zero on the OPEN dataset. The given image x^0 is highlighted with a red box, and the text prompt y is shown under the set of frames. There are a total of 128 video frames ($M = 127$), and the synthesized results for every 14 frames are presented.

centrally-cropped and resized to 256×256.

4.3. Result Analysis

Ablation Study. We conduct ablation study of different sampling strategies on MUG. As shown in Tab. 1 and Fig. 4, compared with generating using randomly sampled Gaussian noise, initializing the input noise with DDPM inversion is important for generating temporally continuous videos, improving all of the metrics dramatically. For MUG, increasing the DDIM sampling steps from 10 to 50 does not enhance the video quality but requires more inference time. Thus, we choose 10-step DDIM as the default setting on MUG. As shown in Fig. 4 and Tab. 1, adding resampling helps preserve identity details (*e.g.*, hairstyle and facial appearance), resulting in lower FVD scores. Increasing resampling steps from 2 to 4 further improves FVD scores.

Effect of Real/Synthesized Starting Frames. We also explore the effect of video generation starting with real or synthesized frames on UCF101. We initially use the first frame of the real videos to generate videos with our TI2V-Zero, termed TI2V-Zero-Real. Additionally, we utilize the backbone model ModelScopeT2V [60] to generate synthetic videos using the text inputs of UCF101. We then employ TI2V-Zero to create videos from the first frame of the generated fake videos, denoted as TI2V-Zero-Fake. As shown in Tab. 2, [TI2V-Zero-Fake *vs.* ModelScopeT2V] can achieve better FVD scores than [TI2V-Zero-Real *vs.* Real Videos]. The reason may be that frames generated by ModelScopeT2V can be considered as *in-distribution* data since TI2V-Zero is built upon it. We also compare the output video distribution of TI2V-Zero-Fake and ModelScopeT2V with real videos in Tab. 2. Though starting from the same synthesized frames, TI2V-Zero-Fake can generate more realistic videos than the backbone model.

Comparison with SOTA Model. We compare our proposed TI2V-Zero with DynamiCrafter [67] in Tab. 3 and Fig. 5. From Fig. 5, one can find that DynamiCrafter struggles to preserve details from the given image, and the motion of its generated videos is also less diverse. Note that DynamiCrafter requires additional fine-tuning to enable TI2V generation. In contrast, without using any fine-tuning or introducing external modules, our proposed TI2V-Zero can precisely start with the given image and output more visually-pleasing results, thus achieving much better FVD scores on both MUG and UCF101 datasets in Tab. 3. The comparison between our TI2V-Zero models with and without using resampling in Fig. 5 and Tab. 3 also demonstrates the effectiveness of using resampling, which can help maintain identity and background details.

Extension to Other Applications. TI2V-Zero can also be extended to other tasks as long as we can construct $\mathbf{s}_0$ with K images at the beginning. These images can be obtained either from ground truth videos or by applying the repeating operation. Then we can slide $\mathbf{s}_0$ when generating the subsequent frames. We have applied TI2V-Zero in video infilling (see the last row in Fig. 3), video prediction (see Supplementary Materials), and long video generation (see Fig. 6). As shown in Fig. 6, when generating a 128-frame video on the OPEN dataset, our method can preserve the mountain shape in the background, even at the 71st frame (frame $\hat{x}^{70}$). The generated video examples and additional experimental results are in our Supplementary Materials.

5. Conclusion

In this paper, we propose a zero-shot text-conditioned image-to-video framework, TI2V-Zero, to generate videos by modulating the sampling process of a pretrained video diffusion model without any optimization or fine-tuning. Comprehensive experiments show that TI2V-Zero can achieve promising performance on multiple datasets.

While showing impressive potential, our proposed TI2V-Zero still has some limitations. First, as TI2V-Zero relies on a pretrained T2V diffusion model, the generation quality of TI2V-Zero is constrained by the capabilities and limitations of the pretrained T2V model. We plan to extend our method to more powerful video diffusion foundation models in the future. Second, our method sometimes generates videos that are blurry or contain flickering artifacts. One possible solution is to apply post-processing methods such as blind video deflickering [30] or image/video deblurring [51] to enhance the quality of final output videos or the newly synthesized frame in each generation. Finally, compared with GAN and standard video diffusion models, our approach is considerably slower because it requires running the entire diffusion process for each frame generation. We will investigate some faster sampling methods [29, 32] to reduce generation time.

References

[1] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4. IEEE, 2010. 2, 6

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 2

[3] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 2

[4] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Bjorn Ommer. Understanding object dynamics for interactive image-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5171–5181, 2021. 2

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 7

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019. 2

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2

[9] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 4

[10] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G Derpanis, and Bjorn Ommer. Stochastic image-to-video synthesis using cinns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3742–3753, 2021. 2

[11] 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. 3

[12] Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10681–10692, 2023. 1, 2

[13] 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. 2

[14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[16] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2, 7

[17] 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. 2

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 6

[21] 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, 7

[22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2, 7

[23] 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. 2

[24] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022. 6

[25] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18219–18228, 2022. 1, 2, 7

[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola,

Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 7

[27] 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

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[29] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021. 8

[30] Chenyang Lei, Xuanchi Ren, Zhaoxiang Zhang, and Qifeng Chen. Blind video deflickering by neural filtering with a flawed atlas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10439–10448, 2023. 8

[31] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[32] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 8

[33] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2, 6

[34] Aniruddha Mahapatra and Kuldeep Kulkarni. Controllable animation of fluid elements in still images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2022. 2

[35] 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. *arXiv preprint arXiv:2108.01073*, 2021. 2

[36] 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.

[37] Nithin Gopalakrishnan Nair, Anoop Cherian, Suhas Lohit, Ye Wang, Toshiaki Koike-Akino, Vishal M Patel, and Tim K Marks. Steered diffusion: A generalized framework for plug-and-play conditional image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20850–20860, 2023. 2

[38] Haomiao Ni, Yihao Liu, Sharon X Huang, and Yuan Xue. Cross-identity video motion retargeting with joint transformation and synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 412–422, 2023. 2

[39] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023. 1, 2, 7

[40] Haomiao Ni, Jiachen Liu, Yuan Xue, and Sharon X Huang. 3d-aware talking-head video motion transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4954–4964, 2024. 2

[41] 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. 3

[42] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[43] OpenAI. Openai: Introducing chatgpt. URL https://openai.com/blog/chatgpt, 2022. 6

[44] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022. 2

[45] 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. 2

[46] 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. 4

[47] 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, 4, 7

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[49] 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. 2

[50] 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

[51] Siddhant Sahu, Manoj Kumar Lenka, and Pankaj Kumar Sa. Blind deblurring using deep learning: A survey. *arXiv preprint arXiv:1907.10128*, 2019. 8

[52] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m:

Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[53] 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

[54] 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, 3

[55] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3

[56] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 6

[57] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[58] 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. 3

[59] 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. 1

[60] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1, 2, 3, 7, 8

[61] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 2

[62] 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. 2

[63] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 1, 2

[64] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022. 2

[65] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 2

[66] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2

[67] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 1, 2, 7, 8

[68] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2

[69] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 2

[70] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 1, 2

Supplementary Materials for TI2V-Zero: Zero-Shot Image Conditioning for Text-to-Video Diffusion Models

Haomiao Ni[1*] Bernhard Egger[2] Suhas Lohit[3] Anoop Cherian[3] Ye Wang[3]
Toshiaki Koike-Akino[3] Sharon X. Huang[1] Tim K. Marks[3]
[1]The Pennsylvania State University, USA [2]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[3]Mitsubishi Electric Research Laboratories (MERL), USA
[1]{hfn5052, suh972}@psu.edu [2]bernhard.egger@fau.de [3]{slohit,cherian,yewang,koike,tmarks}@merl.com

https://merl.com/demos/TI2V-Zero

A. Dataset Details

We conduct extensive experiments on three diverse datasets, including facial expression dataset MUG, action recognition dataset UCF101, and our self-created dataset OPEN. Here we present comprehensive details about these datasets.

For the MUG dataset, we randomly select 5 male and 5 female subjects from the available 52 individuals, and 4 expressions from the provided 7 expression classes. Detailed information about selected subjects and corresponding expression labels are presented in Tab. 1. To convert expression class labels to text prompts for input, we use the following templates: "`A woman with the expression of slight {label} on her face.`" for female subjects, and "`A man with the expression of slight {label} on his face.`" for male subjects. Considering that the average original video length on MUG is about 72 frames, we uniformly sample 16 frames from most of the videos to create the real videos. For videos with more than 80 frames, we sample the videos every 5 frames until we obtain 16 frames to form the real videos.

For the UCF101 dataset, we initially randomly select some action classes from the provided 101 classes. Subsequently, we identify and choose 10 action classes where both ModelScopeT2V and VideoCrafter1 are able to synthesize promising videos. Table 2 shows the details of selected action class labels and their corresponding text prompts. For each action class, we simply choose the first 10 subjects. Given that the average original video length on the UCF101 dataset is approximately 200 frames, we sample the videos every 10 frames until 16 frames are obtained to form the real videos. For videos containing less than 160 frames, we uniformly sample 16 frames.

For the OPEN dataset, we first employ ChatGPT 3.5[1] to generate 10 text prompts by inputting the query "Could you randomly generate 10 text prompts for testing text-to-video models?". We list these 10 text prompts in Tab. 3. Then we use Stable Diffusion 1.5 with the model ID `dreamlike-photoreal-2.0`[2] to generate 100 images for each of 10 text prompts, resulting in a total of 1,000 images as starting frames.

Male ID	007, 010, 013, 014, 020
Female ID	001, 002, 006, 046, 048
Expression	Anger, Happiness, Sadness, Surprise

Table 1. Details of selected subjects and expression classes on the MUG dataset.

Action Class	Text Prompt
ApplyEyeMakeup	"A person is applying eye makeup."
BabyCrawling	"A baby is crawling."
BreastStroke	"A person is performing breaststroke."
Drumming	"A person is drumming."
HorseRiding	"A person is riding horse."
Kayaking	"A person is kayaking."
MilitaryParade	"Military parade."
PlayingGuitar	"A person is playing guitar."
Surfing	"A person is surfing."
ShavingBeard	"A person is shaving beard."

Table 2. Details of selected action class labels and corresponding text prompts on the UCF101 dataset.

B. Additional Experimental Results

More Prior Work Comparisons. In Tab. 4, we conduct additional experiments to compare our proposed model with the open-domain TI2V model VideoComposer [1] on MUG and UCF101 datasets, where our TI2V-Zero achieves supe-

*Work done during an internship at MERL.
[1]https://openai.com/blog/chatgpt
[2]https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0

1	“A mesmerizing display of the northern lights in the Arctic.”
2	“A bustling street market in Marrakech with colorful textiles and spices.”
3	“A futuristic cityscape with holographic advertisements and flying cars.”
4	“A romantic gondola ride through the canals of Venice at sunset.”
5	“A group of friends on a road trip, singing along to their favorite songs.”
6	“A serene mountain cabin covered in a fresh blanket of snow.”
7	“A thrilling skateboarder performing tricks in a skate park.”
8	“A bustling night market in Bangkok with street food vendors and live music.”
9	“A high-speed bullet train racing through a scenic countryside.”
10	“A group of explorers uncovering the mysteries of an ancient temple in the jungle.”

Table 3. The 10 text prompts used in the OPEN dataset.

Figure 1. Examples of generated video frames in video prediction task conditioning on different numbers of given images. The 1st, 6th, 11th, and 16th frames of each output video are shown in each column. Each generated video has 16 frames with a resolution of 256×256. `1 image`, `4 images`, `8 images` indicate the use of the first 1, 4, and 8 real video frames in the ground truth video to predict the next 15, 12, and 8 frames, respectively.

rior performance.

Extension to Video Prediction Task. We have presented the results of video infilling and long video generation in the main paper. In Fig. 1 and our supplementary videos, we show the application of our proposed TI2V-Zero to the video prediction task. Specifically, we conduct experiments using the first 1, 4, and 8 real video frames from the ground truth videos to generate 16-frame videos, i.e., synthesize the subsequent 15, 12, and 8 frames, respectively. As illustrated in Fig. 1, when only 1 image is provided, the woman in the generated video applies the powder brush to the eye differently from the real video. With 4 images, the woman in the synthesized video applies the brush to the same eye as in the ground truth video, but it is still hard to maintain the same appearance of the brush as real video. When extending to 8 frames, the model can synthesize a video that is consistent with the given real video.

Model	MUG			UCF101	
	FVD↓	sFVD↓	tFVD↓	FVD↓	tFVD↓
VideoComposer [1]	1899.08	2294.60±482.45	2050.69±116.92	633.32	1606.13±355.87
TI2V-Zero (Ours)	**180.09**	**267.17±74.72**	**252.77±39.02**	**477.19**	**1306.75±271.82**

Table 4. Quantitative comparison between VideoComposer and TI2V-Zero (w/ Resample) for TI2V generation.

DDIM	Resample	Time (s)
10	0	5.46
50	0	24.86
10	2	14.90
10	4	24.70

Table 5. The average inference time for generating one frame using our proposed TI2V-Zero under different sampling settings. The terms `DDIM` and `Resample` represent the number of steps of using DDIM sampling and resampling.

Inference Time and GPU Usage. In Tab. 5, we report the average inference time of generating one frame with our proposed TI2V-Zero under different sampling settings, when using a batch size of one on a Quadro RTX 6000 GPU. The GPU usage for each setting is 9,885 MB. With the same GPU, the baseline DynamiCrafter takes about 155 seconds to generate a 16-frame video using their default settings.

C. Discussion with Concurrent Work

A concurrent work to ours, AnimateZero [2], also adopts a similar repeating operation. However, we are different in several aspects. In our framework, when computing temporal attention outputs, the sources of keys are derived either from the given image or previously synthesized images, whereas AnimateZero relies on keys from the given image or noise. Moreover, AnimateZero shares keys and values from spatial self-attention of the first frame across the other frames; this may make it hard to generate large motions and novel scenes, as the content is constrained to the informa-

tion available in the first frame. In contrast, our framework demonstrates the ability to generate promising videos containing intricate motions with input images of various styles across a wide variety of scenes.

D. Information about Example Videos

We include eight MP4 files of example video clips generated by our proposed method in the Supplementary materials.

- **mug.mp4** includes the video clips generated by the state-of-the-art model DynamiCrafter and our proposed TI2V-Zero for 4 expressions of 4 subjects from the MUG dataset.
- **ucf.mp4** contains the synthesized video clips produced by DynamiCrafter and our TI2V-Zero for action classes from the UCF101 dataset.
- **open.mp4** contains the generated video clips using DynamiCrafter and our TI2V-Zero for 10 text prompts from the OPEN dataset.
- **ablation.mp4** compares the generated video clips under different sampling strategies using our proposed TI2V-Zero on the MUG dataset. The terms `Inversion`, `DDIM`, and `Resample` denote the application of DDPM-based inversion, the steps using DDIM sampling, and the iteration number using resampling, respectively.
- **long_video.mp4** displays one example video clip showing the application of our proposed TI2V-Zero to generate a 128-frame long video.
- **prediction.mp4** shows one example video clip illustrating the application of our proposed TI2V-Zero to video prediction task, conditioning on different numbers of given images.
- **motivation.mp4** shows the video clips generated by the replacing-based baseline approach and our proposed TI2V-Zero for different video tasks (corresponding to Fig. 3 in our main paper).
- **intricate.mp4** shows the video clips generated with intricate text and image inputs, including two (16-frame) videos and one long (64-frame) video. Each first-frame image in the video clips was generated by Stable Diffusion 1.5.

References

[1] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 1, 2

[2] Jiwen Yu, Xiaodong Cun, Chenyang Qi, Yong Zhang, Xintao Wang, Ying Shan, and Jian Zhang. Animatezero: Video diffusion models are zero-shot image animators. *arXiv preprint arXiv:2312.03793*, 2023. 2