Unified Text-Image-to-Video Generation: A Training-Free Approach to Flexible Visual Conditioning

Anonymous authors
Paper under double-blind review

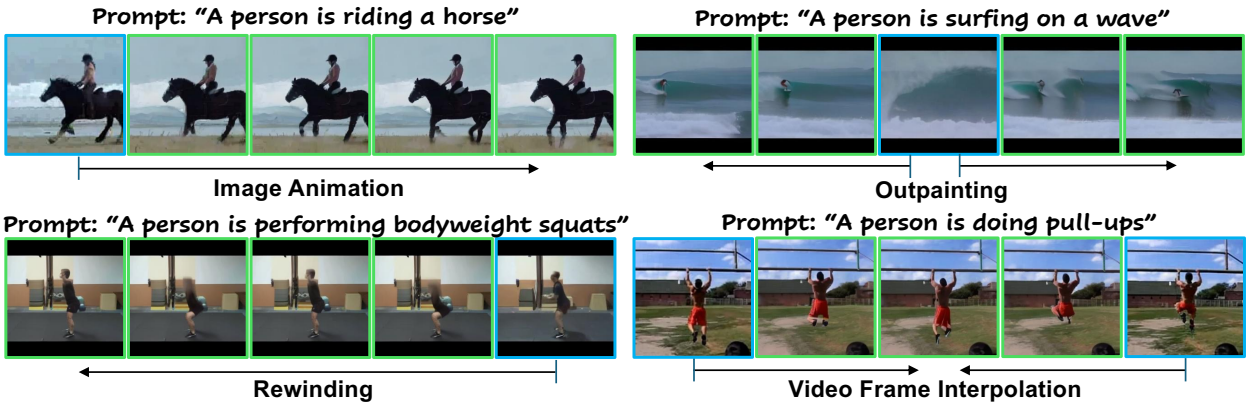

Figure 1: We propose **FlexTI2V**, a novel training-free approach that can add flexible image conditioning to off-the-shelf text-to-video foundation models. We are able to frame an *arbitrary number* of images at *arbitrary positions* in the synthetic video with vivid motion and smooth transitions. In this figure, images with blue edges are condition images, and images with green edges are generated video frames.

Abstract

Text-image-to-video (TI2V) generation is a critical problem for controllable video generation using both semantic and visual conditions. Most existing methods typically add visual conditions to text-to-video (T2V) foundation models by finetuning, which is costly in resources and only limited to a few pre-defined conditioning settings. To tackle these constraints, we introduce a unified formulation for TI2V generation with flexible visual conditioning. Furthermore, we propose an innovative training-free approach, dubbed **FlexTI2V**, that can condition T2V foundation models on an arbitrary amount of images at arbitrary positions. Specifically, we firstly invert the condition images to noisy representation in a latent space. Then, in the denoising process of T2V models, our method uses a novel random patch swapping strategy to incorporate visual features into video representations through local image patches. To balance creativity and fidelity, we use a dynamic control mechanism to adjust the strength of visual conditioning to each video frame. Extensive experiments validate that our method surpasses previous training-free image conditioning methods by a notable margin. Our method can also generalize to both UNet-based and transformer-based architectures across various image conditioning settings.

1 Introduction

In recent years, diffusion models (Ho et al., 2020) have shown excellent performance in text-to-video (T2V) generation (Ho et al., 2022; He et al., 2022b; Singer et al., 2023; Hong et al., 2023; Kong et al., 2024; Blattmann et al., 2023b; Wang et al., 2025a; Yang et al., 2025; Wang et al., 2023), serving as video foundation models for

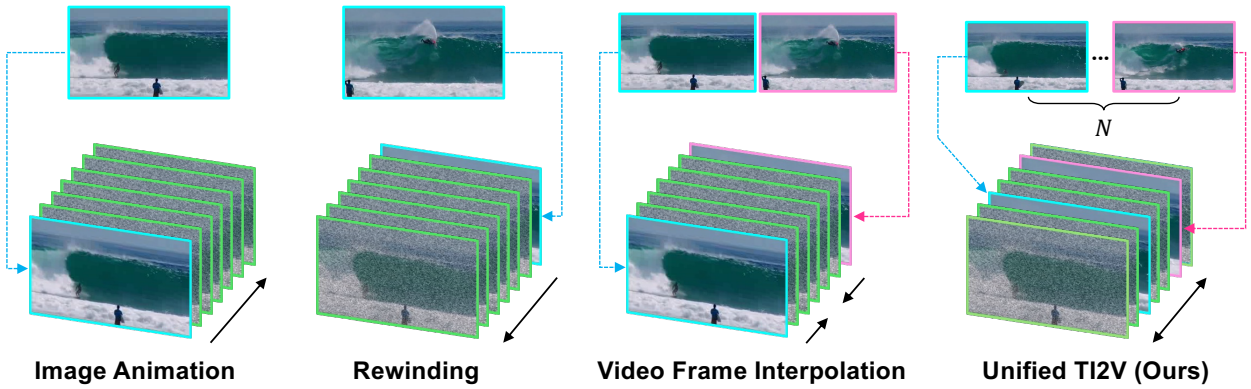

Figure 2: **Comparison with classic TI2V tasks.** Our task requires video generation conditioned on any number of images at any positions in the output video, which unifies existing classic TI2V tasks. The images with blue and pink edges are condition images, and images with green edges are generated video frames.

various downstream video generation tasks. Given the vagueness and ambiguity of language (Lai et al., 2025; 2024), recent work has investigated synthesizing videos conditioned on RGB images for precise controllability on the appearance of people, objects and background (Blattmann et al., 2023a; Li et al., 2024; Chen et al., 2024b), which is as known as image-to-video (I2V) generation. In these models, RGB images can be used as conditions alone (Blattmann et al., 2023a; Li et al., 2024; Wang et al., 2022; Mahapatra & Kulkarni, 2022), or incorporated with textual prompts for accurate guidance (Zhang et al., 2023; Guo et al., 2023; Xing et al., 2024; Huang et al., 2025). For clarity, we explicitly refer to the latter as text-image-to-video (TI2V) generation following (Huang et al., 2025; Ni et al., 2024).

To include extra image conditioning in video generation, the most widely-used approach is finetuning pre-trained T2V model with paired texts, images and videos. However, the prohibitive demands of high-quality data and computing resources, and the risks of involving unexpected bias from the data during finetuning impede the development of TI2V models. Moreover, as shown in Figure 2, off-the-shelf TI2V models can only apply the image conditioning setting pre-defined during training (*e.g.*, image animation (Guo et al., 2023; Zhang et al., 2024; Chen et al., 2024a) and video frame interpolation (Sim et al., 2021; Reda et al., 2022)), lacking sufficient flexibility in the number and positions of condition images. Any changes in the conditioning setting require re-training of the entire model. To avoid the computing overhead and undesired bias of finetuning, training-free image conditioning approaches are promising solutions to incorporate flexible visual control into off-the-shelf T2V models.

Prior to our work, recent studies have been conducted on training-free controllable video generation with various conditions (Kara et al., 2024; Khachatryan et al., 2023; Zhou et al., 2024; Hou et al., 2025; Liang et al., 2025a;b). For training-free image conditioning, the previous methods either synthesize only one video frame in one denoising pass, making it inefficient in inference (Ni et al., 2024), or fill in the missing video frames between two bounded images solely based on visual cues, losing semantic control of generated frames in between (Feng et al., 2024). Similar to finetuned TI2V models, they are also only capable of following pre-defined conditioning settings, and still short of flexibility for image conditioning. These inevitable deficiencies make the existing training-free methods suboptimal for TI2V generation problem, thus demanding more efforts and novel solutions.

In this work, we formally define a generic TI2V generation problem with high flexibility in the amounts and positions of condition images (see Figure 1 and Figure 2), unifying the existing TI2V tasks such as image animation, rewinding and video frame interpolation. To address aforementioned issues, we introduce ***FlexTI2V***, an innovative flexible training-free approach that can condition an off-the-shelf text-to-video diffusion model on an arbitrary number of images at arbitrary positions in the synthetic video. Specifically, we first reverse the images to noisy representations at each denoising step. To harmonize the difference in denoising paths of T2V models and image conditioning, we propose a novel *random patch swapping* (RPS)

method to incorporate visual features of input images into each video frame, by swapping a random portion of patches between video frames and condition images. We also use a dynamic adjustment strategy to control the percentage of swapped patches and denoising steps for each frame to prevent frozen videos. The extensive experiments suggest that our approach outperforms the previous training-free image conditioning methods for video diffusion models. Overall, our main contributions are summarized as follows:

- We formally define a generic text-image-to-video generation problem, unifying the various settings of existing TI2V tasks.
- We introduce FlexTI2V, a novel training-free approach to incorporate flexible image conditioning into pretrained text-to-video diffusion models. The key innovation is a random patch swapping method to incorporate visual guidance into the denoising process in a plug-and-play manner.
- The experiments indicate that our approach notably surpasses the existing training-free image conditioning methods. Our method also generalizes seamlessly to both UNet and transformer architectures.

2 Related Work

Video Generation Conditioned on Texts and Images Text-to-video (T2V) generation has gained impressive advancement since the emergence of diffusion models (Ho et al., 2022; He et al., 2022b; Hong et al., 2022; Blattmann et al., 2023b; Kong et al., 2024; Yang et al., 2025; Wang et al., 2025a; 2023). Given the ambiguity of language in describing visual details, RGB images are exploited as the condition for video generation. Image conditioning can be used alone for image-to-video (I2V) generation (Blattmann et al., 2021; Dorkenwald et al., 2021; Wang et al., 2022; Mahapatra & Kulkarni, 2022; Blattmann et al., 2023a; Li et al., 2024), or integrated with textual prompts for text-image-to-video (TI2V) generation (Ni et al., 2023; Zhang et al., 2023; Jiang et al., 2024; Chen et al., 2024a; Wang et al., 2025b; Huang et al., 2025). Here, we mainly discuss previous TI2V methods. Hu et al. (2022) propose TI2V task for the first time, and develop a transformer-based architecture to achieve decent controllability and diversity in generated videos. Guo et al. (2023) develop a practical framework, AnimateDiff, to train a model-agnostic plug-and-play motion module that can turn any personalized text-to-image diffusion model into an image animator. Xing et al. (2024) introduce a novel model – DynamiCrafter, which can animate both human bodies and objects. Fu et al. (2023) propose a unified video completion method that can tackle image animation, rewinding and infilling using a single model. Most prior methods require training some layers or the entire model to incorporate image conditioning or motion over time. As a result, existing models can only handle TI2V tasks pre-defined in training, such as animation and infilling. In this paper, we introduce a generic TI2V problem that unifies previous TI2V tasks. Our proposed method is able to condition T2V diffusion models on an arbitrary number of images at arbitrary positions of the output videos without finetuning.

Video Frame Interpolation Video frame interpolation has been extensively studied in recent years, with the model architectures spanning across convolutional networks (Niklaus et al., 2017; Liu et al., 2020; Reda et al., 2022), transformers (Fu et al., 2023) and diffusion models (Feng et al., 2024; Repo, 2024; 2025). Liu et al. (2017) propose a convolutional network to synthesize video frames by flowing pixel values from bounded images. Xu et al. (2019) point out that most interpolation methods assume a uniform motion between two consecutive images, thus resulting in oversimplified linear models for interpolation. To generate complex motion flows, They propose a quadratic interpolation method to model acceleration information in videos. Previous work mainly aims at finding a plausible path of transition from the first video frame to the last one. How to go beyond the two bounded images remains understudied. In this work, we work on the generic TI2V problem including both classic video frame interpolation and frame synthesis beyond the bound of start and end images.

Training-free Methods for Diffusion Foundation Models The high computing overhead and large amounts of training data of finetune the foundation models are inevitable bottlenecks for downstream tasks. Recently, the development of training-free approaches provide an excellent trade-off between performance and cost (Hertz et al., 2022; Mokady et al., 2023; Tumanyan et al., 2023; Liu et al., 2024; Khachatryan et al., 2023; Kara et al., 2024; Ni et al., 2024; Titov et al., 2024; Feng et al., 2024; Zhou et al., 2024; Hou et al., 2025; Liang et al., 2025a;b; Hsiao et al., 2025). Hertz et al. (2022) introduce prompt-to-prompt method to edit synthetic images by manipulating self-attention maps, which is further extended to real photo editing using

null-text inversion (Mokady et al., 2023). Kara et al. (2024) adapt a pretrained text-guided image editing model for video editing tasks by random noise shuffling. In terms of image-conditioned video generation, the work most related to our method is TI2V-Zero (Ni et al., 2024). They duplicate the condition image and synthesize only one video frame in a single denoising process, which enforces the pretrained temporal layer to leverage visual features from the image. However, this method suffers from slow inference speed and compounding error. Feng et al. (2024) convert an image-to-video model to a bounded-image interpolation model by merging the two separate denoising paths conditioned on the bounded images. Hou et al. (2025) propose to map a single image to 3D point cloud and then synthesize new videos from other viewpoints. Existing training-free TI2V methods usually focus on a specific problem setting. We make the first attempt to develop a training-free method for various image conditioning settings, with high flexibility in image amounts and positions.

3 Preliminaries

Latent Diffusion Models Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) are initially proposed in RGB space, which is computationally expensive because of the high resolution. As a solution, latent diffusion models (LDM) (Rombach et al., 2022) project RGB images or videos into a low-dimensional latent space with a visual encoder $\mathcal{E}$. The denoising process is conducted in the latent space to generate the latent representation of desired images or videos, which are decoded to RGB space by a decoder $\mathcal{D}$. To achieve controllable generation, the conditions are also embedded into the same space, and typically merged into the model through cross-attention layers.

Denoising Diffusion Implicit Models (DDIM) DDIM (Song et al., 2020) basically include two stages – diffusion stage and denoising stage. Given a clean representation $\boldsymbol{z}_0$ in the latent space, Gaussian noise is added to the $\boldsymbol{z}_0$ step by step in the diffusion stage. The single-step sample $\boldsymbol{z}_t$ at step t also follows a Gaussian distribution, $\boldsymbol{z}_t \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\,\boldsymbol{z}_0,\ (1-\bar{\alpha}_t)\boldsymbol{I})$. The closed-form equation is

$$\boldsymbol{z}_t = \sqrt{\bar{\alpha}_t} \cdot \boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t} \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\alpha_t = 1-\beta_t$. β_t is the noise schedule. In denoising stage, DDPM build the denoising process as a Markov chain, and estimates the distribution of $\boldsymbol{z}_{t-1}$ from $\boldsymbol{z}_t$. The denoising process is slow due to a large amount of steps (typically 1000 steps). To accelerate sampling, DDIM define a more general non-Markovian process to estimate the Gaussian distribution of $\boldsymbol{z}_{t-1}$ conditioned on $\boldsymbol{z}_t$ and $\boldsymbol{z}_0$, thus reducing sampling steps significantly. After substitution of $\boldsymbol{z}_0$ with an expression of $\boldsymbol{z}_t$, DDIM sampling is formulated as

$$\boldsymbol{z}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}} \left(\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t} \cdot \boldsymbol{\varepsilon}_\theta(\boldsymbol{z}_t, t)\right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \cdot \boldsymbol{\varepsilon}_\theta(\boldsymbol{z}_t, t) + \sigma_t \cdot \boldsymbol{\epsilon}', \tag{2}$$

where $\boldsymbol{\epsilon}' \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. $\boldsymbol{\varepsilon}_\theta$ is a neural network trained to predict the noise added to $\boldsymbol{z}_t$. σ_t is a parameter that controls the stochasticity of the sampling process. Setting $\sigma_t = 0$ yields a deterministic path that can reduce sampling steps with minor loss of generation quality (Song et al., 2020). Then Equation 2 can be simplified to a function of $\boldsymbol{z}_t$ and t, *i.e.*, $\boldsymbol{z}_{t-1} = \boldsymbol{\mu}_\theta(\boldsymbol{z}_t, t)$.

4 FlexTI2V

The setting of the proposed generic text-image-to-video (TI2V) generation problem is illustrated in Figure 2. The input includes a textual prompt y, N images $\boldsymbol{x} = \{x^{(n)}\}_{n=0}^{N-1}$ as visual conditions, and the expected position index $\boldsymbol{p} = \{p^{(n)}\}_{n=0}^{N-1}$ of each image in the output video. The goal is to synthesize a video consisting of M frames $\boldsymbol{v} = \{v^{(m)}\}_{m=0}^{M-1}$ with each condition image at the expected position (*i.e.*, $v^{\left(p^{(n)}\right)} = x^{(n)}$ where $p^{(n)} \in [0, M-1]$). The output video is supposed to show a reasonable transition across condition images following the textual prompt. Our problem setting unifies many existing TI2V tasks in one formulation, such as image animation ($N=1$, $\boldsymbol{p}=\{0\}$), video rewinding ($N=1$, $\boldsymbol{p}=\{M-1\}$) and video frame interpolation ($N=2$, $\boldsymbol{p}=\{0, M-1\}$).

The major challenge of training-free TI2V generation is harmonizing the distinct denoising directions controlled by texts and images solely relying on frozen model parameters. The experiments of MAE (He et al.,

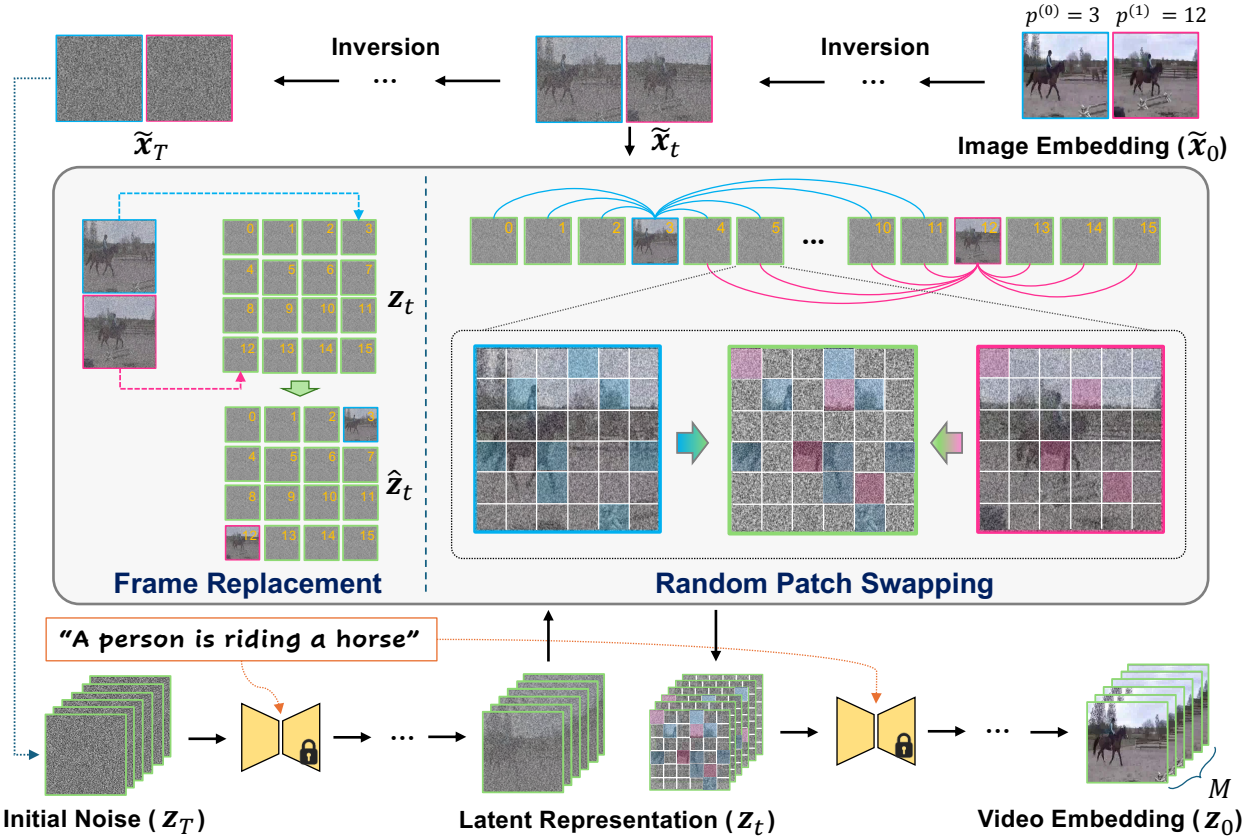

Figure 3: **Overview of the proposed FlexTI2V approach.** We invert the condition image embedding to noisy representation $\tilde{\boldsymbol{x}}_t$ at each step. The final noise $\tilde{\boldsymbol{x}}_T$ is reused as initialization for video synthesis. At step t, we directly replace the video frames with images at the desired positions. Then, for each video frame, we randomly swap a portion of patches with bounded condition images based on the relative distance between the frame and images. Though we show a special case of using two condition images in this figure, our method can naturally extend to any number of images at any positions. Note that all operations of our method occur in the latent space. We visualize RGB images and frames on the latent representations simply for intuitive understanding.

2022a) validate that global image information can be carried by and recovered from a portion of local patches. Inspired by this finding, the key insight of our *FlexTI2V* method is contextualizing each synthetic video frame with visual conditioning, by randomly swapping a portion of patches between video frames and condition images in the latent space at early denoising steps. To balance the fidelity to condition images and creativity in motion generation, we use a dynamic control strategy for the percentage of swapped patches and denoising steps for each video frame, based on their distance to the condition images. This strategy also introduces position information of condition images into the denoising process, thus making the diffusion model capable of handling multiple condition images at any desired positions.

4.1 Condition Image Inversion and Replacement

Image Inversion The overview of our method is illustrated in Figure 3. First of all, we encode condition images $\boldsymbol{x}$ into latent embedding $\tilde{\boldsymbol{x}}$ using the pretrained encoder $\mathcal{E}$, *e.g.*, $\tilde{\boldsymbol{x}} = \mathcal{E}(\boldsymbol{x}) = \{\tilde{x}^{(n)}\}_{n=0}^{N-1}$. To fuse the visual condition with T2V denoising process, we invert the clean latent image embedding into noisy representations at each denoising step. We denote the representation at step $t = 0$ as $\tilde{\boldsymbol{x}}_0 = \tilde{\boldsymbol{x}}$ for consistent notations in diffusion models. The widely-used DDIM inversion (Mokady et al., 2023; Titov et al., 2024) is

not applicable to our problem because the video diffusion model can not estimate the noise accurately in a single image without temporal information. Instead, we use the straightforward DDPM-based inversion as an alternative following Meng et al. (2022) and Ni et al. (2024), which directly adds Gaussian noise to $\tilde{\boldsymbol{x}}_0$ like diffusion process, $\tilde{\boldsymbol{x}}_t \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\,\tilde{\boldsymbol{x}}_0,\ (1-\bar{\alpha}_t)\boldsymbol{I}\right)$. In denoising process, T2V diffusion models typically initialize the noise by drawing a random sample from Gaussian distribution (Ho et al., 2020). Recent work has validated reusing noise inverted from images preserves the global structure in image editing (Tumanyan et al., 2023; Mokady et al., 2023; Liu et al., 2024), and leads to consistent video frames in video generation (Ni et al., 2024). Hence, we duplicate the inverted noise $\tilde{\boldsymbol{x}}_T$ at the final step, and stack them along the temporal dimension as the initial $\boldsymbol{z}_T$ for video denoising.

Frame Replacement As shown in Figure 3, the representation of the M video frames at step t is denoted as $\boldsymbol{z}_t = \{z_t^{(m)}\}_{m=0}^{M-1}$. We leverage the pretrained temporal modeling layer in the T2V diffusion model to propagate visual features of condition images to all video frames. Inspired by Ni et al. (2024), we replace the video frame embeddings at the conditioning position $\boldsymbol{p} = \{p^{(n)}\}_{n=0}^{N-1}$ with inverted image embeddings, which is formulated as

$$\texttt{FrameReplace}\left(\tilde{\boldsymbol{x}}_t, \boldsymbol{z}_t, \boldsymbol{p}\right) = \left\{z_t^{(0)}, z_t^{(1)}, \cdots, z_t^{\left(p^{(n)}-1\right)}, \tilde{x}_t^{\left(p^{(n)}\right)}, z_t^{\left(p^{(n)}+1\right)}, \cdots, z_t^{(M-2)}, z_t^{(M-1)}\right\}, \tag{3}$$

where $n \in [0, N-1]$. We refer to the resulting video frame embedding as $\hat{\boldsymbol{z}}_t$.

4.2 Random Patch Swapping

Due to the limited number of condition images compared with the long video frame sequence, the features of condition images may be neglected in the temporal layers. The generated video thus still deviates from the visual conditioning. To tackle this challenge, we introduce a novel *random patch swapping* (RPS) method to explicitly incorporate visual features into each video frame by swapping a random portion of patches between each pair of video frame embedding $\hat{z}_t^{(m)}$ and condition image embedding $\tilde{x}_t^{(n)}$ (see Figure 3 for visualization). Specifically, we pre-define a percentage, $P(m, n, t)$, of swapped patches between m-th video frame and n-th condition image at step t. Then we generate a random binary mask $\mathcal{M}_t^{(m,n)}$ composed of zeros and ones having the same shape as $\hat{z}_t^{(m)}$ and $\tilde{x}_t^{(n)}$, *i.e.*, $\mathcal{M}_t^{(m,n)}$, $\hat{z}_t^{(m)}$, $\tilde{x}_t^{(n)} \in \mathbb{R}^{C\times H\times W}$. The number of ones in the mask account for $P(m, n, t)$ of the total elements, so we have $\sum_{c,i,j} \mathcal{M}_t^{(m,n)}[c, i, j]/(C \times H \times W) = P(m, n, t)$. Note that the values are identical across the channel dimension at each spatial location (i, j), so we omit the channel c in following paragraphs for simplicity.

Random patch swapping is essentially replacing the element of m-th video frame at position (i, j) with the element of n-th condition image at the same position. Each element in the latent space corresponds to a patch in RGB space depending on the encoder compression ratio. Then RPS can be formulated as updating the values in $\hat{z}_t^{(m)}$ following the equation below:

$$\hat{z}_t^{(m)}[i,j] = \mathcal{M}_t^{(m,n)}[i,j] \cdot \tilde{x}_t^{(n)}[i,j] + \left(1 - \mathcal{M}_t^{(m,n)}[i,j]\right) \cdot \hat{z}_t^{(m)}[i,j]. \tag{4}$$

In this way, the visual cues from condition images are delivered to video frames through swapped patches, and then harmonized with video features by the pretrained diffusion foundation models. Each video frame only swaps patches with the two bounded condition images (the closest condition images before and after it). If all condition images are on the same side of a video frame, this frame only swaps patches with the single nearest condition image. We find that a large percentage of swapped patches may include overly strong spatial alignment prior, while a small percentage contains insufficient visual features for effective guidance. To make a trade-off, we dynamically reduce the percentage and steps of RPS when the video frame is far from the position of condition image. The percentage $P(m, n, t)$ is written as

$$P(m,n,t) = \begin{cases} P_0 - \delta_1 \cdot |m - p^{(n)}|, & T - \tilde{t} < t \le T, \\ 0, & 0 \le t \le T - \tilde{t}, \end{cases} \quad \text{where} \quad \tilde{t} = t_0 - \delta_2 \cdot |m - p^{(n)}|. \tag{5}$$

P_0 and t_0 are initial swapping percentage and steps, and δ_1, δ_2 are decreasing strides. Our dynamic control strategy is based on the intuitive hypothesis that video frames around condition images are also perceptually

Table 1: **Quantitative comparison with previous methods.** "User" denotes the user study results. ↓ means a lower score indicates better performance. The orange row refers to our method.

Methods	User	FVD↓	tFVD↓	CLIP-T	CLIP-F	CLIP-I
Image Animation						
DynamiCrafter (Xing et al., 2024)	26.20	210.77	832.38	29.86	89.10	80.95
TI2V-Zero (Ni et al., 2024)	28.40	189.49	586.54	29.84	88.01	82.29
FlexTI2V	**45.40**	**125.49**	**558.29**	**30.97**	**94.40**	**84.38**
Rewinding						
TI2V-Zero (Ni et al., 2024)	39.80	199.01	611.60	29.77	89.21	81.02
FlexTI2V	**60.20**	**134.98**	**596.96**	**30.95**	**94.58**	**85.11**
Outpainting						
TI2V-Zero (Ni et al., 2024)	38.60	216.80	641.07	28.80	89.01	85.01
FlexTI2V	**61.40**	**167.46**	**632.71**	**30.75**	**95.89**	**88.34**
Video Frame Interpolation						
DynamiCrafter (Xing et al., 2024)	16.20	310.74	731.90	29.69	84.26	78.68
TRF (Feng et al., 2024)	32.20	247.88	631.54	29.47	91.99	83.35
FlexTI2V	**51.60**	**98.41**	**406.30**	**30.92**	**92.34**	**86.88**

similar, thus demanding more visual cues by patch swapping. Likewise, for frames far from condition images, we reduce constraints of visual conditioning accordingly to encourage more creativity in motion generation. We only swap patches in the first $\tilde{t}$ denoising steps, and let the model to harmonize and smooth artifacts in the remaining $T - \tilde{t}$ steps by setting $P(m, n, t) = 0$. Additionally, This method also introduces position information of images into a T2V model, which makes our approach applicable to any number of images at any positions. The entire algorithm formulation is shown in Appendix A.

5 Experiments

5.1 Dataset and Metrics

Dataset We run experiments on the widely-used UCF-101 dataset (Soomro et al., 2012) following Ho et al. (2022), Fu et al. (2023) and Ni et al. (2024), which is composed of 13,320 videos of 101 action categories. We select 42 action classes including complex actions with high dynamics (*e.g.*, surfing) and small motion with still background (*e.g.*, typing). For each class, we randomly select 25 instances (totaling 1,050 videos) as the test set. We manually rewrite the textual prompts for each action. We also run experiments on one more synthetic dataset to consolidate our conclusion, which is shown in the supplementary.

Metrics We use three CLIP-based metrics including (1) CLIP frame-text similarity (CLIP-T) measuring the fidelity of generated frames to textual prompts, (2) CLIP frame similarity (CLIP-F) measuring the consistency across all video frames, and (3) CLIP image similarity (CLIP-I) measuring the agreement between condition images and video frames at other positions. We also use (4) Frechet Video Distance (FVD) (Unterthiner et al., 2018) to assess the overall video quality, its variant (5) text-conditioned FVD (tFVD) (Ni et al., 2024) to assess the video-text alignment, and (6) user study for human evaluation.

Implementation Details For a fair comparison with Ni et al. (2024), we implement our FlexTI2V method on ModelScopeT2V (Wang et al., 2023), which shows great robustness to out-domain T2V generation. We set the number of video frames as $M = 16$, and total denoising steps $T = 20$ with deterministic DDIM sampling (*i.e.*, $\sigma_t = 0$ in Equation 2). The initial percentage and steps of swapping are set as $P_0 = 0.3$ and $t_0 = 10$. The decreasing strides are $\delta_1 = 5 \times 10^{-3}$ and $\delta_2 = 0.3$. Our method can be run on 1 NVIDIA H100 GPU. Please refer to Appendix C.2 for more implementation details.

Table 2: **Comparison of model efficiency in inference.** The orange column refers to our method. Our method shows great advantage in inference speed compared with other baseline models.

	DynamiCrafter (Xing et al., 2024)	TI2V-Zero (Ni et al., 2024)	TRF (Feng et al., 2024)	**FlexTI2V**
Inference Time (s)	9.45	275.14	60.05	**2.71**

5.2 Quantitative Comparison

We compare with prior approaches quantitatively under three classic settings: (1) image animation – one condition image as the first video frame ($N = 0$, $\boldsymbol{p} = \{0\}$), (2) rewinding – one condition image as the last video frame ($N = 0$, $\boldsymbol{p} = \{15\}$), and (3) video frame interpolation – two condition images as the first and last frames ($N = 2$, $\boldsymbol{p} = \{0, 15\}$). We also introduce a new setting, video frame outpainting – one condition image as the middle video frame ($N = 1$, $\boldsymbol{p} = \{7\}$). We mainly compare with two state-of-the-art *training-free* methods – TI2V-Zero (Ni et al., 2024) and time reversal fusion (TRF) (Feng et al., 2024). TI2V-Zero is initially designed only for image animation task. We adapt it to outpainting and rewinding tasks by adjusting the order of frame synthesis. Moreover, we also include DynamiCrafter (Xing et al., 2024) as a *training-based* baseline for a thorough comparison. More details of baseline implementation are elaborated in the supplementary. The results are shown in Table 1.

Single-Image Conditioning For image animation, outpainting and rewinding tasks, our method outperforms DynamiCrafter and TI2V-Zero by a notable margin in FVD, which suggests an overall higher quality of videos synthesized by our method. In fine-grained evaluation metrics, our FlexTI2V method surpasses the best baseline by 28.25/8.36/14.64 in tFVD and 1.11%/1.95%/1.18% in CLIP-T for the three tasks, indicating the best prompt-following capability. In addition, the two baselines have low scores in CLIP-F and CLIP-I, indicating the inconsistency across generated video frames and low fidelity to the condition images. The reason could be that DynamiCrafter is a training-based model which still suffers from the bottleneck of generalization to out-of-domain data. Though TI2V-Zero is a training-free method, it synthesizes the future video frames autoregressively, resulting in compounding errors and cumulative deviations from visual conditions, which can also be observed in visualization in Section 5.3. In contrast, our method generates all video frames simultaneously in a single denoising process, thus achieving the best performance in CLIP-F and CLIP-I. The prominent improvement (5.30%/6.88%/5.37% and 2.09%/3.33%/4.09%) also validates the effectiveness of our random patch swapping strategy for various image conditioning settings.

Multi-Image Conditioning In terms of video frame interpolation, DynamiCrafter still lags behind TRF and FlexTI2V due to domain shift. In contrast, TRF shows better performance for this task (measured by FVD) relying on the strong generalization capability of stable video diffusion (Blattmann et al., 2023a), which further validates the advantage of training-tree method in controllable video generation. Though TRF achieves a high score in CLIP-F and CLIP-I, it still has a limited performance in video-text alignment (tFVD and CLIP-T) due to lack of semantic control by textual prompts. In contrast, our method outperforms both competitors in all the metrics. The significant improvement further validates the flexibility of our method to multi-image conditioning.

User Study Besides automatic metrics. We also conduct user study for a thorough evaluation. We sample 100 instances from our test set for each task, with each instance evaluated by five raters. Each rater is asked to select the video with best alignment with the textual prompt and condition images. The human preference percentage of each method is presented in Table 1. FlexTI2V is selected as the best in the highest percentage of samples compared with other competitors. The results suggest that the videos synthesized by our method align better with human's subjective perception.

Inference Efficiency We evaluate the average inference time of our method and the baseline models for generating a single 16-frame video in Table 2. The test is conducted on one NVIDIA H100 GPU. TI2V-Zero synthesizes video frames one by one, and relies on multiple resample steps to improve the performance. The inference time thus is significantly longer than other approaches. TRF conducts denoising processes conditioned on two bounded images separately and then merges the features at each step, resulting in longer

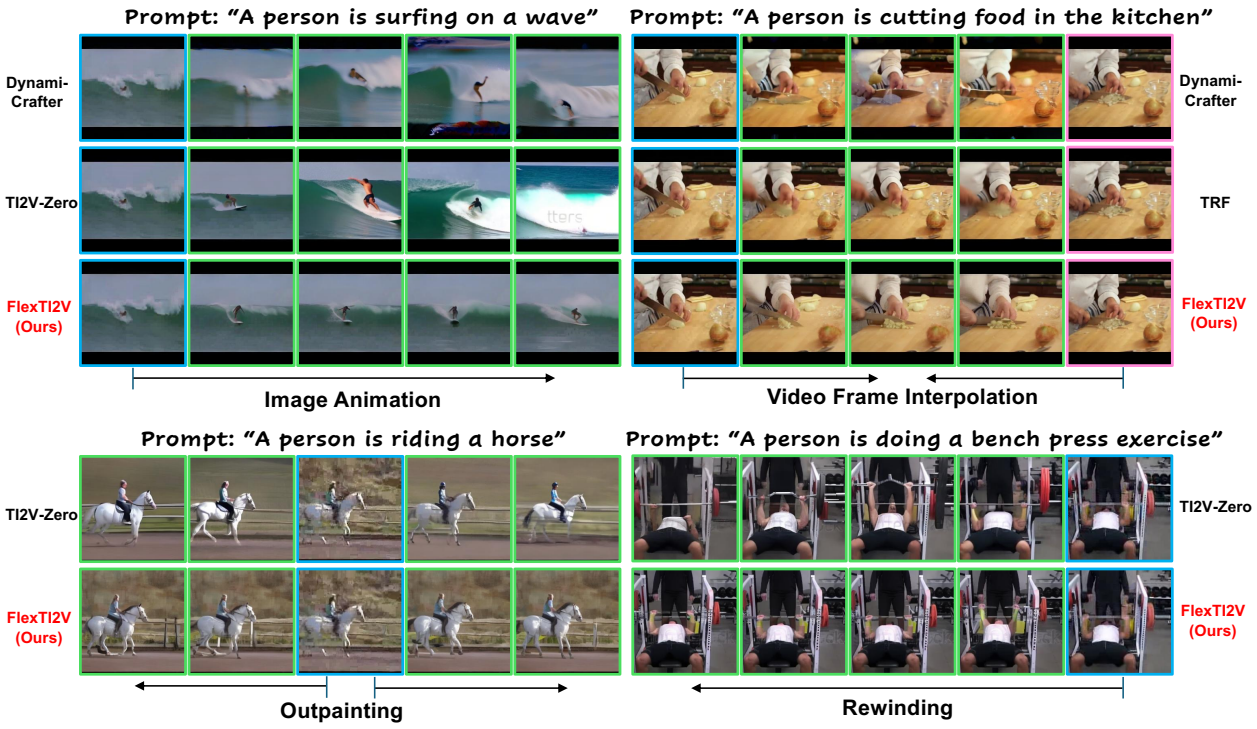

Figure 4: **Comparison with prior methods.** The images with blue and pink edges are condition images for each setting, and images with green edges are generated video frames. Our approach synthesizes videos with higher frame consistency and fidelity to condition images than other baseline models in various settings.

inference time than DynamiCrafter. In contrast, our method requires only 2.71 seconds for 20 DDIM denosing steps, which is much lower than DynamiCrafter and TRF. The great gap validates the high efficiency of our method.

5.3 Visualization for Qualitative Comparison

Besides quantitative evaluation, we also visualize the videos synthesized by our method and other baselines for a qualitative comparison. As illustrated in Figure 4, we observe some distortion in the videos generated by DynamiCrafter (*e.g.*, the person in image animation and the knife in video frame interpolation), which may be the explanation for the low performance in Table 1. TRF is good at preserving visual features of the two bounded images. However, the video frames synthesized by TRF are blurry around the knife. We speculate that TRF is initially proposed for smooth motion generation and camera viewpoint transition. It may be weak in modeling object state change, such as the onion being cut into small pieces in this example. TI2V-Zero can generate videos with large motion because of the autoregressive generation mechanism. However, the cumulative mistakes weaken the controllability of visual conditions, leading to a low consistency in the frames far from the condition image. In contrast, our method can follow the textual prompts and visual conditions faithfully, and synthesize smooth motions in all TI2V settings. More examples of our method are illustrated in Appendix B.3.

5.4 Generalization to Different Foundation Models

We use ModelScopeT2V (Wang et al., 2023) as the foundation model in previous experiments simply for a fair comparison with prior methods. To validate the generalization of our method, we also implement our method to another SOTA T2V foundation model – Wan2.1-T2V (Wang et al., 2025a) using transformer-based architectures. We directly compare with Wan2.1-I2V and Wan2.1-FLF2V on image animation and

Table 3: **Comparison with Wan2.1 model family.** We compare our method with Wan2.1-I2V on image animation and compared with Wan2.1-FLF2V on video frame interpolation. ↓ means a lower score in this metric indicates a better performance. The orange column refers to our method.

Methods	FVD↓	CLIP-T	CLIP-I
Image Animation			
Wan2.1-I2V (Wang et al., 2025a)	141.22	31.01	82.33
Wan2.1-T2V + Our Method	**101.74**	**30.98**	**85.87**
Video Frame Interpolation			
Wan2.1-FLF2V (Repo, 2025)	112.22	31.47	85.13
Wan2.1-T2V + Our Method	**90.82**	**31.50**	**86.04**

Figure 5: **Generated videos of our method implemented to Wan2.1-T2V.** The images with blue edges are condition images and images with green edges are generated video frames. Please refer to supplementary for corresponding videos and extra demos.

video frame interpolation separately. The two specialized models are finetuned from the base Wan2.1-T2V model for the two specific tasks, which makes it an apple-to-apple comparison with our proposed approach.

As shown in Table 3, our method outperforms the two specialized models on the two tasks. The possible reason is that it's still challenging for training-based methods to generalize to out-of-distribution data (*e.g.*, different aspect ratios and resolutions). This also shows the necessity of training-free methods like our work. We demonstrate videos synthesized by our method in Figure 5. Implemented on top of the modern architecture, our method generates videos with high fidelity and smooth motions, showing robust generalization to different T2V models.

Table 4: **Ablation study of components in our method.** ↓ means a lower score in this metric indicates a better performance. The orange row refers to our complete method. Each component is removed from the complete method *separately*.

	Animation		Outpainting		Rewinding		Interpolation	
	FVD↓	CLIP-I	FVD↓	CLIP-I	FVD↓	CLIP-I	FVD↓	CLIP-I
FlexTI2V	**125.49**	**84.38**	**167.46**	**88.34**	**134.98**	**85.11**	**98.41**	**86.88**
w/o Frame Replace	130.50	83.90	174.02	87.80	137.20	84.50	104.11	86.51
w/o RPS	440.28	64.41	510.69	70.02	430.21	66.20	425.86	68.23

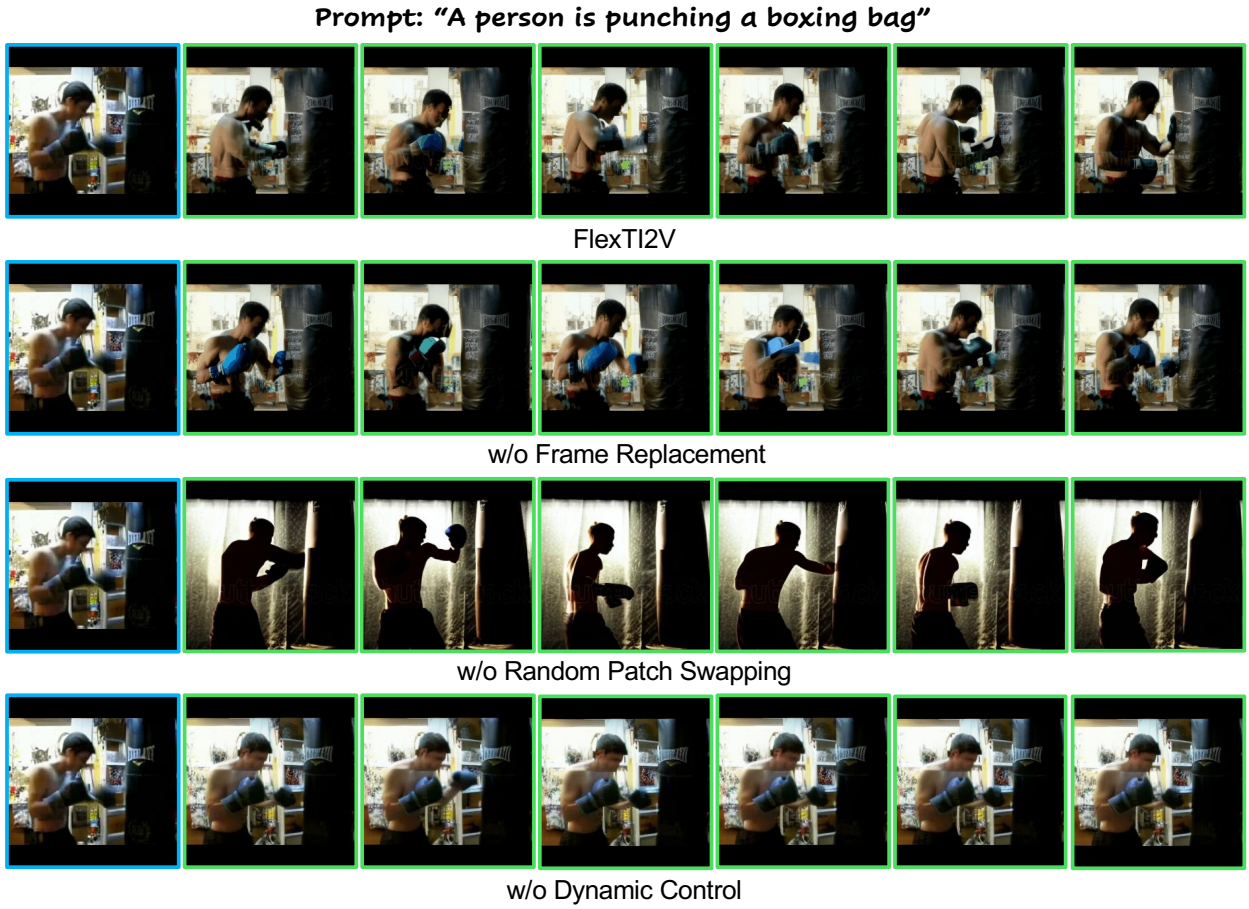

Figure 6: **Demonstration of ablation study.** The images with blue edges are condition images and images with green edges are generated video frames. More explanations are provided in Section 5.5.

5.5 Ablation Study

In our method, there are two components (frame replacement and random patch swapping) to incorporate visual conditioning into T2V diffusion models. We evaluate the contribution of each component by ablating them from our method *separately*. The results are presented in Table 4. After removing frame replacement, the performance lags behind the full model by a small margin in all tasks. The output video is still decent with minor deviation from the condition image (second row in Figure 6). The result indicates that random patch swapping (RPS) can work alone to effectively merge visual guidance into T2V diffusion process. It also validates our hypothesis that sufficient visual features can be transferred through a portion of local patches at each denoising step for video generation. On the contrary, if we remove RPS and only implement

Table 5: **Hyperparameter analysis.** We change only one hyperparameter in each subtable and keep the other hyperparameters fixed. ↓ means a lower score indicates better performance. The orange columns refer to the hyperparameters we finally adopt.

P_0	0.2	**0.3**	0.4
FVD↓	156.64	**125.49**	147.11

t_0	7	**10**	13
FVD↓	136.65	**125.49**	134.96

δ_1	4×10^{-3}	$\mathbf{5 \times 10^{-3}}$	6×10^{-3}	7×10^{-3}
FVD↓	133.79	**125.49**	137.46	151.70

δ_2	0.2	**0.3**	0.4	0.5
FVD↓	133.79	**125.49**	137.46	151.70

frame replacement strategy for extra image conditioning, a significant performance drop is observed across all metrics and tasks. Though frame replacement introduces image features into the temporal layers, the model estimates the holistic noise added to all video frames (16 frames in our experimental setting) as an entirety. Hence, the condition image features are likely to be ignored because of the small number of images against the large number of video frames. We observe that the videos generated without RPS greatly deviate from the condition images (the third row in Figure 6), thus resulting in a remarkable decrease in the performance. In the last ablation experiment, we remove the dynamic control strategy from our method and use a fixed percentage and steps for random patch swapping. The results are illustrated in the last row of Figure 6. The motion in the generated video frames is trivial without dynamic control. The reason is because the constraints of visual conditioning are overly strong for all frames. The result validates the necessity of the dynamic control strategy.

5.6 Hyperparameter Analysis

In the proposed RPS approach, there are four important hyperparameters – initial swapping percentage P_0, initial number of swapping steps t_0 and the decreasing strides δ_1, δ_2 for swapping percentage and steps. The optimal hyperparameters are selected empirically based on the analysis in Table 5. If we use a large δ_1 and δ_2, or small P_0 and t_0, the image conditioning is weaker than the optimal hyperparameters so that the generated video can not accurately follow the images. Likewise, if δ_1 and δ_2 are small, or P_0 and t_0 are large, the image conditioning is too strong so that the output videos are nearly frozen. To achieve an optimal balance, we select the best hyperparameter combination based on the experiments. We also find the optimal combination is very consistent across different datasets, so we use it in all the experiments.

6 Conclusion

In this paper, we introduce a generic text-image-to-video (TI2V) problem that unifies existing TI2V tasks, which requires the model be able to synthesize smooth videos conditioned on an arbitrary number of images at any expected positions. To solve this problem, we propose a novel training-free method, FlexTI2V, that can introduce flexible image conditioning to pretrained text-to-video diffusion models. The key component of our method is random patch swapping strategy, which merges visual features of condition images into video denoising process by swapping a random portion of local patches. The extensive experiments indicate a notable improvement of FlexTI2V over prior methods, as well as the flexibility of our model. Our work is an important step for flexible and low-cost controllable video generation, which provides an alternative way to adapt pretrained foundation models to specific downstream tasks, thus democratizing the access to customized video generation tools.

References

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*, pp. 5171–5181, 2021.

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

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*, pp. 22563–22575, 2023b.

Xi Chen, Zhiheng Liu, Mengting Chen, Yutong Feng, Yu Liu, Yujun Shen, and Hengshuang Zhao. Livephoto: Real image animation with text-guided motion control. In *European Conference on Computer Vision*, pp. 475–491. Springer, 2024a.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. SEINE: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2024b. URL `https://openreview.net/forum?id=FNq3nIvP4F`.

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*, pp. 3742–3753, 2021.

Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *European Conference on Computer Vision*, pp. 378–395. Springer, 2024.

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*, pp. 10681–10692, 2023.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022a.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022b.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

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.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. In *International Conference on Learning Representations*, 2025.

Teng-Fang Hsiao, Bo-Kai Ruan, Yi-Lun Wu, Tzu-Ling Lin, and Hong-Han Shuai. Tf-ti2i: Training-free text-and-image-to-image generation via multi-modal implicit-context learning in text-to-image models. *arXiv preprint arXiv:2503.15283*, 2025.

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*, pp. 18219–18228, 2022.

Haoyang Huang, Guoqing Ma, Nan Duan, Xing Chen, Changyi Wan, Ranchen Ming, Tianyu Wang, Bo Wang, Zhiying Lu, Aojie Li, et al. Step-video-ti2v technical report: A state-of-the-art text-driven image-to-video generation model. *arXiv preprint arXiv:2503.11251*, 2025.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6689–6700, 2024.

Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6507–6516, 2024.

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

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Bolin Lai, Xiaoliang Dai, Lawrence Chen, Guan Pang, James M Rehg, and Miao Liu. Lego: L earning ego centric action frame generation via visual instruction tuning. In *European Conference on Computer Vision*, pp. 135–155. Springer, 2024.

Bolin Lai, Felix Juefei-Xu, Miao Liu, Xiaoliang Dai, Nikhil Mehta, Chenguang Zhu, Zeyi Huang, James M Rehg, Sangmin Lee, Ning Zhang, and Tong Xiao. Unleashing in-context learning of autoregressive models for few-shot image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24142–24153, 2024.

Feng Liang, Akio Kodaira, Chenfeng Xu, Masayoshi Tomizuka, Kurt Keutzer, and Diana Marculescu. Looking backward: Streaming video-to-video translation with feature banks. In *International Conference on Learning Representations*, 2025a.

Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, et al. Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7817–7826, 2024.

Yihao Liu, Liangbin Xie, Li Siyao, Wenxiu Sun, Yu Qiao, and Chao Dong. Enhanced quadratic video interpolation. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 41–56. Springer, 2020.

Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE international conference on computer vision*, pp. 4463–4471, 2017.

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*, pp. 3667–3676, 2022.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

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*, pp. 18444–18455, 2023.

Haomiao Ni, Bernhard Egger, Suhas Lohit, Anoop Cherian, Ye Wang, Toshiaki Koike-Akino, Sharon X Huang, and Tim K Marks. Ti2v-zero: Zero-shot image conditioning for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9015–9025, 2024.

Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 670–679, 2017.

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Frame interpolation for large motion. In *European Conference on Computer Vision*. Springer, 2022.

GitHub Repo. Cogvideox-interpolation: Keyframe interpolation with cogvideox. https://github.com/feizc/CogvideX-Interpolation, 2024.

GitHub Repo. Wan2.1-flf2v. https://github.com/Wan-Video/Wan2.1, 2025.

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*, pp. 10684–10695, 2022.

Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. Xvfi: extreme video frame interpolation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14489–14498, 2021.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=nJfylDvgzlq`.

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

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.

Vadim Titov, Madina Khalmatova, Alexandra Ivanova, Dmitry Vetrov, and Aibek Alanov. Guide-and-rescale: Self-guidance mechanism for effective tuning-free real image editing. In *European Conference on Computer Vision*, pp. 235–251. Springer, 2024.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1921–1930, 2023.

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.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025a.

Cong Wang, Jiaxi Gu, Panwen Hu, Yuanfan Guo, Xiao Dong, Hang Xu, and Xiaodan Liang. Dreamvideo: High-fidelity image-to-video generation with image retention and text guidance. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

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.

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.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024.

Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

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

Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7747–7756, 2024.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

A Algorithm Formulation

We formulate our FlexTI2V approach in Algorithm 1. The function $\texttt{Repeat}(\tilde{x}_T^{(0)}, M)$ at step 3 duplicates the inverted condition image representation for M times as the initialization for video frame denoising. If there are more than one condition images, we find using different image makes trivial difference in the generation performance. We use the first condition image in our method. The function $\texttt{BoundIndex}(m, \boldsymbol{p})$ at step 8 returns the indices of the *two* closest condition images before and after m-th video frame respectively (*i.e.*, bounded index). If all condition images are before or after this video frame, the function returns only *one* index of the closest image.

Algorithm 1 FlexTI2V

Require: The condition images $\boldsymbol{x}$; The expected positions of each image in the output video $\boldsymbol{p}$; The textual prompt y; A pretrained text-to-video diffusion foundation model, including a paired encoder $\mathcal{E}$ and decoder $\mathcal{D}$ as well as noise estimation network $\boldsymbol{\varepsilon}_\theta$. Hyperparameters including the initial percentage P_0, number of denoising steps t_0, the decreasing strides δ_1 and δ_2, and the amount of video frames M.

Ensure: A generated video of M frames $\boldsymbol{v}$, with each condition image appearing at the expected position, *i.e.*, $v^{(p^{(n)})} = x^{(n)},\ \forall n \in [0, N-1]$.

1: $\tilde{\boldsymbol{x}}_0 \leftarrow \mathcal{E}(\boldsymbol{x})$
2: $\tilde{\boldsymbol{x}}_T \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_T}\ \tilde{\boldsymbol{x}}_0,\ (1-\bar{\alpha}_t)\boldsymbol{I}\right)$
3: $\boldsymbol{z}_T \leftarrow \texttt{Repeat}(\tilde{x}_T^{(0)},\ M)$ ▷ *Noise Initialization*
4: **for** $t \in \{T, T-1, \cdots, 2, 1\}$ **do**
5: $\quad\tilde{\boldsymbol{x}}_t \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\ \tilde{\boldsymbol{x}}_0,\ (1-\bar{\alpha}_t)\boldsymbol{I}\right)$ ▷ *Image Inversion*
6: $\quad\hat{\boldsymbol{z}}_t \leftarrow \texttt{FrameReplace}(\tilde{\boldsymbol{x}}_t,\ \boldsymbol{z}_t,\ \boldsymbol{p})$ ▷ *Frame Replacement, Equation 3 in Main Paper*
7: $\quad$**for** $m \in \{0, 1 \cdots, M-1\}$ **do**
8: $\quad\quad\boldsymbol{n}_{bound} \leftarrow \texttt{BoundIndex}(m, \boldsymbol{p})$ ▷ *Indices of Bounded Condition Image*
9: $\quad\quad$**for** $n \in \boldsymbol{n}_{bound}$ **do**
10: $\quad\quad\quad\tilde{t} \leftarrow t_0 - \delta_2 \cdot |m - p^{(n)}|$ ▷ *Dynamic Control in Swapping Steps*
11: $\quad\quad\quad$**if** $T - \tilde{t} < t \leq T$ **then**
12: $\quad\quad\quad\quad P(m, n, t) \leftarrow P_0 - \delta_1 \cdot |m - p^{(n)}|$ ▷ *Dynamic Control in Swapping Percentage*
13: $\quad\quad\quad$**else**
14: $\quad\quad\quad\quad P(m, n, t) \leftarrow 0$
15: $\quad\quad\quad$**end if**
16: $\quad\quad\quad$Generate random binary mask $\mathcal{M}_t^{(m,n)}$ with element "1" accounting for $P(m, n, t)$.
17: $\quad\quad\quad$**for** $i \in \{0, 1, \cdots, H-1\}, j \in \{0, 1, \cdots, W-1\}$ **do** ▷ *Random Patch Swapping*
18: $\quad\quad\quad\quad\hat{z}_t^{(m)}[i,j] \leftarrow \mathcal{M}_t^{(m,n)}[i,j] \cdot \tilde{x}_t^{(n)}[i,j] + \left(1 - \mathcal{M}_t^{(m,n)}[i,j]\right) \cdot \hat{z}_t^{(m)}[i,j]$
19: $\quad\quad\quad$**end for**
20: $\quad\quad$**end for**
21: $\quad$**end for**
22: $\quad\boldsymbol{z}_t \leftarrow \hat{\boldsymbol{z}}_t$
23: $\quad\boldsymbol{z}_{t-1} \leftarrow \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t} \cdot \boldsymbol{\varepsilon}_\theta(\boldsymbol{z}_t, t)\right) + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \boldsymbol{\varepsilon}_\theta(\boldsymbol{z}_t, t)$ ▷ *Deterministic Sampling*
24: **end for**
25: $\boldsymbol{v} \leftarrow \mathcal{D}(\boldsymbol{z}_0)$
26: **return** $\boldsymbol{v}$

Table 6: **Results on the synthetic dataset.** The orange column refers to our method. Refer to Section B.1 for more descriptions.

Methods	Animation		Interpolation	
	CLIP-T	CLIP-I	CLIP-T	CLIP-I
DynamiCrafter (Xing et al., 2024)	24.33	79.53	25.44	75.50
TRF (Feng et al., 2024)	-	-	25.92	81.38
TI2V-Zero (Ni et al., 2024)	24.87	81.76	-	-
FlexTI2V	**26.83**	**82.31**	**27.29**	**83.54**

B Additional Experiment Resutls

B.1 Quantitative Results on Synthetic Dataset

In our experiments, we follow the previous work (Ni et al., 2024) to conduct evaluation on UCF-101 for a fair comparison. To make the testing data more diverse, we utilize the data curation pipeline of (Ni et al., 2024) to synthesize an OPEN dataset by Stable Video Diffusion using 10 prompts:

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

Different from UCF-101, the 10 prompts are mainly about non-human motions and object state changes. The synthetic data also guarantees the base T2V model is not trained on it. We generate 22 videos for each prompt, totaling 220 testing samples.

As shown in Table 6, our method still prominently outperforms all the baseline approaches in CLIP-T (text-video alignment) and CLIP-I (image-video similarity). The results validate that our method is also superior in modeling different types of motions and videos in the TI2V task.

B.2 Generated Videos without Humans

The videos in UCF-101 are mainly about human actions. To assess the generalization performance of our approach, we supplement extra examples of FlexTI2V synthesizing videos without humans in Figure 7. Our approach seamlessly generalizes to the motion of objects showing strong generalization capability.

B.3 Additional Demonstration of Our Method

We illustrate more videos synthesized by our method in classic TI2V generation settings (see Figure 9) and new settings (see Figure 10). Our approach is able to condition the pretrained T2V diffusion foundation model on any number of images with flexible positions.

Figure 7: **Additional examples of our method on videos without humans.** The images with blue edges are condition images and images with green edges are generated video frames. Please see Section B.2 for more details.

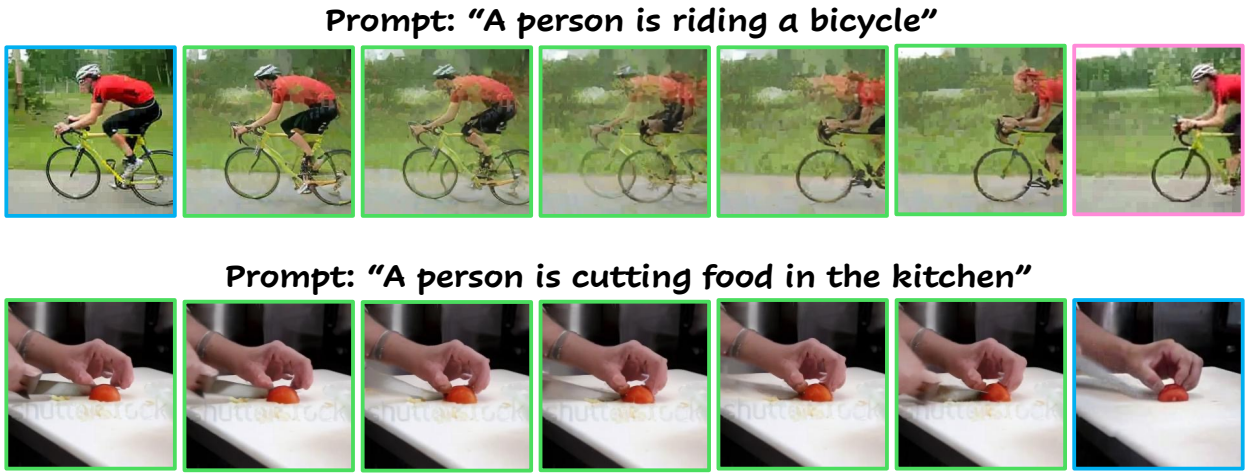

Figure 8: **Typical failure cases of our approach.** The images with blue and pink edges are condition images and images with green edges are generated video frames. Please refer to Section B.4 for more analysis.

B.4 Failure Cases

We demonstrate some failure cases of our method in Figure 8 to show the boundary of the capability. We find it is challenging for our method to synthesize smooth camera viewpoint change in output videos (first example). If there is a big difference in viewpoints of two condition images, our method is very likely to treat the two images as irrelevant images and generates a hard transition without a reasonable motion. In addition, we also observe watermarks in a few output videos (second example). This is an inherent bias of the foundation model, which is inherited by our method. Hence, the performance of our method is still limited by the pretrained T2V foundation model.

Figure 9: Additional demonstration of videos synthesized by FlexTI2V in classic TI2V settings. The images with blue and pink edges are condition imags, and images with green edges are generated video frames. N and $\boldsymbol{p}$ are the number and positions of condition images. Our method generates videos with 16 frames in the experiments.

C Implementation Details

C.1 Establishment of Test Set

Ni et al. (2024) select 10 classes in UCF-101 which are not diverse enough for a reliable evaluation. The textual prompts used in their work are also oversimplified. In this work, we further extend the test set to 42 classes with manually annotated textual prompt for each action. For each class, there are videos of 25 subjects performing the same action. For diversity, we randomly select one video of each subject in each action class, totaling 1050 test data samples. The prompts are listed in Table 7.

C.2 More Details of Our Method

In the proposed method, the model also conditions on textual prompts. We adopt classifier-free guidance in the inference with guidance scale as 9.0. We use a deterministic DDIM sampling strategy ($\sigma_t = 0$ in main paper Equation 2) that can accelerate generation by reducing denoising steps (Song et al., 2020). We generate videos with $M = 16$ frames in all experiments for a fair comparison with previous methods. Note

Figure 10: Additional demonstration of videos synthesized by FlexTI2V in new TI2V settings. The images with blue, pink and orange edges are condition imags, and images with green edges are generated video frames. N and $\boldsymbol{p}$ are the number and positions of condition images. Our method generates videos with 16 frames in the experiments.

that our method can naturally extend to videos with more frames. Our method is implemented on a single NVIDIA H100 GPU.

Table 7: Prompt for each selected action class in UCF-101. More details of the test set is elaborated in Section C.1.

Class	Prompt
ApplyEyeMakeup	"A person is applying eye makeup"
ApplyLipstick	"A person is applying lip stick"
BabyCrawling	"A baby is crawling"
BenchPress	"A person is doing a bench press exercise"
Biking	"A person is riding a bicycle"
BlowDryHair	"A person is blow drying their hair"
BodyWeightSquats	"A person is performing bodyweight squats"
BoxingPunchingBag	"A person is punching a boxing bag"
BrushingTeeth	"A person is brushing their teeth"
CuttingInKitchen	"A person is cutting food in the kitchen"
Drumming	"A person is playing drums"
Fencing	"Two people are participating in fencing"
Haircut	"A person is giving someone a haircut"
HandstandPushups	"A person is performing handstand pushups"
HeadMassage	"A person is doing a head massage for another one"
HorseRace	"People are participating in a horse race"
HorseRiding	"A person is riding a horse"
JumpRope	"A person is jumping rope"
Kayaking	"A person is kayaking in water"
Knitting	"A person is knitting"
Mixing	"A person is mixing ingredients"
MoppingFloor	"A person is mopping a floor"
PlayingCello	"A person is playing a cello"
PlayingFlute	"A person is playing a flute"
PlayingGuitar	"A person is playing a guitar"
PlayingPiano	"A person is playing a piano"
PlayingViolin	"A person is playing a violin"
PullUps	"A person is doing pull-ups"
PushUps	"A person is performing push-ups"
Rafting	"People are rafting in a river"
RockClimbingIndoor	"A person is climbing an indoor rock wall"
Rowing	"People are rowing a boat"
ShavingBeard	"A person is shaving their beard"
SkateBoarding	"A person is skateboarding"
Skiing	"A person is skiing down a slope"
Skijet	"A person is riding a jet ski"
SkyDiving	"A person is skydiving"
Surfing	"A person is surfing on a wave"
TrampolineJumping	"A person is jumping on a trampoline"
Typing	"A person is typing on a keyboard"
WalkingWithDog	"A person is walking a dog"
WritingOnBoard	"A person is writing on a board"

C.3 Implementation of Previous Methods

We implement all baseline models using their publicly released codebases following the default hyperparameter settings. To adapt TI2V-Zero (Ni et al., 2024) to rewinding and outpainting tasks, we reverse the order of frame generation for rewinding, and generate the frames before and after the condition image in two directions separately for outpainting. For TRF (Feng et al., 2024), we use the setting of `gym` in their code.

D Limitation and Future Work

Though our approach achieves great performance in many experimental settings compared with previous methods, we still observe some limitations as illustrated in Figure 8. First of all, FlexTI2V is still weak in understanding the transition of camera viewpoints between two condition images. The output video looks like a transition of two slides if there is a big viewpoint change. Second, our method is affected by some flaws in the pretrained foundation models (*e.g.*, watermark) because of the training-free nature. The upperbound of our method is constrained by the T2V model we used. Third, FlexTI2V is able to make the condition images as exact video frames at expected positions. Our method can not synthesize videos following only the subjects or background in the condition images (*e.g.*, personalized video generation).

Based on these limitations, we propose several promising research problems that deserve future investigation.

- Camera motion modeling is an important problem for vivid video generation. We think camera motion between two video frames can be modeled independently and added to existing training-free video generation approaches.
- More interpretation work is necessary to probe the bias of pretrained video generation models as well as how to separate the bias from the learned unbiased knowledge.
- More efforts are needed to extend our method to other image types, such as depth and keypoints.

E Code and Data Release

We will publicly release our code, selected test data and textual prompts to facilitate future research in similar directions after acceptance.

F Video Demonstration

We also provide several accompanying videos for demonstration.

- `presentation.mp4`: presentation of our work with audio narration
- `demo_ablation.mp4`: video demonstration of the ablation study.
- `demo_classic_settings.mp4`: video demonstration of our method using classic TI2V settings (image animation, rewinding, video frame interpolation).
- `demo_new_settings.mp4`: video demonstration of our method using novel TI2V settings.
- `demo_wan.mp4`: video demonstration of our method applied to Wan2.1-T2V.