TI2V-ADAPTER: EFFICIENT ADAPTATION FROM TEXT-TO-VIDEO MODELS TO TEXT-IMAGE-TO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-image-to-video (TI2V) generation aims to synthesize realistic videos from text prompts and reference images. Existing deep-learning-based TI2V methods either require costly training with large-scale video data or rely on training-free inference-time manipulation of pretrained Text-to-video (T2V) models. While training-intensive methods require time-consuming optimization and substantial computational resources, training-free methods often exhibit limited performance because of the lack of learnable reference-image adaptation. These limitations motivate us to develop a TI2V method with low training cost and improved use of reference-image cues. We therefore propose TI2V-Adapter, an efficient adaptation method that introduces lightweight learnable modules to incorporate reference-image information into pretrained T2V models, enabling T2V-to-TI2V adaptation with minimal training cost. Specifically, TI2V-Adapter consists of a Visual Reference Adapter (VRA) that injects reference-image cues at both the input and feature levels of the pretrained T2V pipeline, and a Video Temporal Controller (VTC) that reinforces image information during denoising. Compared with training-free methods, TI2V-Adapter achieves a **15.3%** FVD improvement with only **0.01%** of the training samples and **0.83%** of the trainable parameters required by training-intensive methods. Code is available in the supplementary material.

1 INTRODUCTION

Text-to-video (T2V) generation aims to synthesize realistic and temporally coherent videos from natural language descriptions. However, text prompts alone are often insufficient to precisely specify fine-grained visual details in generated videos. In many practical scenarios, users expect the generated video to not only follow the text description but also preserve the visual characteristics of a given reference image. This motivates Text-image-to-video (TI2V) generation, where a reference image is introduced as an additional visual condition to guide video synthesis, offering more controllable video creation for applications such as artistic creation (Xing et al., 2024a; Guo et al., 2024), entertainment (Che et al., 2025), and data augmentation (Zhou et al., 2025; Hui et al., 2026).

A straightforward solution for TI2V generation is to train image-conditioned video models with large-scale image-text-video pairs, where substantial model parameters are updated to learn how to effectively incorporate reference-image information. For instance, Step-Video-TI2V (Huang et al., 2025) involves **30B trainable parameters** and is further optimized with **5M training samples**. Although such training-intensive methods can achieve strong generation performance, their heavy training costs make them difficult to reproduce and deploy in resource-constrained scenarios.

To reduce the cost of training-intensive methods, training-free TI2V methods directly leverage pretrained T2V backbones for TI2V generation through inference-time manipulation without additional training or extra trainable parameters. For example, TI2V-Zero (Ni et al., 2024) constructs a training-free generation pipeline on top of a pretrained ModelScopeT2V (Wang et al., 2023) model. Specifically, the encoded reference image is repeatedly incorporated into the generation trajectory during denoising without training. Similarly, FlexTI2V (Lai et al., 2025) guides video generation by replacing part of the generated content with information from the reference image during denoising, enabling image conditioning without adding or updating any model parameters.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Although training-free TI2V methods inject reference-image information into pretrained T2V backbones through inference-time manipulation, these backbones are originally trained for text-conditioned generation without dedicated image-conditioning mechanisms. Consequently, they cannot effectively perceive and exploit the injected image cues, leading to insufficient utilization of reference image information and ultimately suboptimal video generation quality.

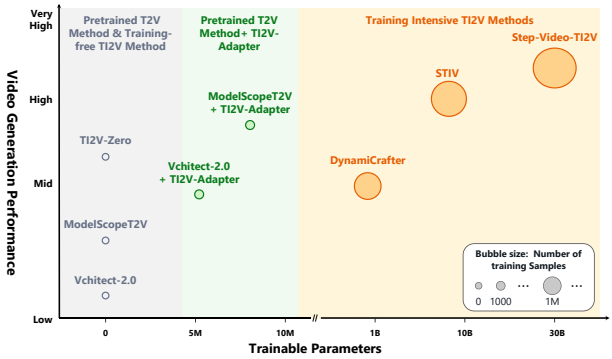

Figure 1: Performance–Efficiency Comparison of Pretrained T2V, Training-Free and Training-Intensive TI2V, and T2V + TI2V-Adapter Methods

These observations raise an important question: ***can pretrained T2V models be effectively adapted to TI2V generation with only a small number of trainable parameters, while better exploiting reference-image information to improve video generation quality?*** We answer this question affirmatively by proposing TI2V-Adapter, which provides an alternative solution to training-free and training-intensive TI2V generation. TI2V-Adapter is a lightweight trainable reference-image module that can be plugged into a pretrained T2V backbone, enabling competitive TI2V performance with limited training data and trainable parameters. Effective T2V-to-TI2V adaptation requires not only injecting reference-image information into the pretrained backbone, but also preventing such information from fading during denoising. To this end, TI2V-Adapter consists of two key components: a Visual Reference Adapter (VRA) and a Video Temporal Controller (VTC). VRA incorporates reference-image information at both the input and feature levels: it injects the reference image into the latent input space via latent fusion for explicit visual guidance, and introduces lightweight trainable parameters into the pretrained T2V backbone for reference-image feature adaptation. Since reference image cues may gradually weaken during denoising, VTC reinforces reference-image information via gated cross-attention as generation proceeds. It is worth noting that TI2V-Adapter can be flexibly adapted to different pretrained T2V backbones, effectively improving their TI2V generation performance with limited trainable parameters and training data, as illustrated in Fig. 1.

In summary, the main contributions of this paper are three-fold:

- We propose TI2V-Adapter, a new parameter-efficient solution that strikes a balance between training-free and training-intensive TI2V generation, enabling pretrained T2V models to effectively exploit reference-image information with minimal training cost.
- We devise two lightweight modules in TI2V-Adapter: a Visual Reference Adapter (VRA) and a Video Temporal Controller (VTC). VRA captures and injects visual cues at both the input and feature levels, while VTC maintains the influence of the reference image.
- Extensive experiments on two benchmark datasets and three diverse backbones demonstrate the effectiveness and architecture-agnostic generalization of TI2V-Adapter. Compared with training-free methods, TI2V-Adapter achieves a **15.3%** FVD improvement with only **0.01%** of the training samples and **0.83%** of the trainable parameters required by training-intensive methods.

2 Related Work

In this section, we briefly review related works on Text-to-video Generation and Text-image-to-video Generation.

2.1 Text-to-video Generation

Text-to-video (T2V) generation aims to synthesize realistic and temporally coherent videos from natural language descriptions. Early attempts explored GAN-based frameworks for translating tex-

tual descriptions into short video clips (Li et al., 2018). With the proliferation of diffusion models (Ho et al., 2020; Rombach et al., 2022), diffusion-based T2V (Ho et al., 2022; Singer et al., 2023; Blattmann et al., 2023) backbones have become increasingly powerful in generating high-fidelity videos. For instance, ModelScopeT2V (Wang et al., 2023) builds a latent diffusion T2V model with a spatio-temporal UNet, enabling efficient video generation from text prompts. More recently, Vchitect-2.0 (Fan et al., 2025) advances this line by introducing a parallel transformer architecture with Multimodal Diffusion Blocks and memory-efficient training, improving the scalability and temporal coherence of T2V generation. These advances indicate that T2V models have learned strong generative priors that can be utilized for TI2V generation when reference images are available.

2.2 Text-image-to-video Generation

Recent Text-image-to-video (TI2V) generation methods extend T2V generation by incorporating reference images as visual conditions, enabling generated videos to follow textual descriptions while preserving the appearance of the reference image. To achieve high generation quality, Step-Video-TI2V and STIV (Huang et al., 2025; Lin et al., 2025) rely on large-scale video training and substantial trainable parameters to learn image-conditioning mechanisms. Despite their compelling generation quality, these methods often require considerable training data, computational resources, and optimization time, making them difficult to reproduce and deploy in resource-constrained scenarios.

In contrast, training-free TI2V methods eliminate additional training by keeping all pretrained T2V parameters frozen and injecting reference-image information solely through inference-time manipulation. For instance, TI2V-Zero (Ni et al., 2024) constructs a training-free generation pipeline on top of ModelScopeT2V (Wang et al., 2023), where the reference image is encoded into latent space and used to replace generated-frame latents during denoising. Moreover, FlexTI2V (Lai et al., 2025) further proposes to inject the reference-image information by swapping local reference-image patches with intermediate video representations during denoising, without additional training. However, since frozen T2V models are originally designed for text-conditioned generation, they lack learnable image branches to perceive and adapt reference-image information. As a result, training-free TI2V methods may insufficiently utilize image cues, leading to suboptimal reference-guided generation performance.

3 Method

3.1 Problem Formulation

Given a text prompt c, a diffusion-based Text-to-video (T2V) neural network generates a video sequence $V = \{x_1, \ldots, x_T\}$ aligned with the textual description (Ho et al., 2020; Nichol & Dhariwal, 2021; Song et al., 2021b) :

$$V = G(z, c; W_\theta), \tag{1}$$

where $z \sim \mathcal{N}(0, I)$ denotes the initial Gaussian noise, and $x_i \in \mathbb{R}^{C \times H \times W}$ denotes the i-th video frame, where C, H, and W represent the number of frame channels, frame height, and frame width, respectively. T is the total number of frames, and $G(\cdot; W_\theta)$ denotes the pretrained T2V model parameterized by W_θ.

In the Text-image-to-video (TI2V) setting, the model further takes a reference image I_{ref} as an additional condition and generates a video sequence that is consistent with the textual description while preserving the visual content of the reference image:

$$V = G(z, c, I_{ref}; W_\phi), \tag{2}$$

where W_ϕ denotes the parameters of the image-conditioned video generation model.

Compared with T2V, TI2V introduces an additional image condition while still benefiting from the text-video alignment and temporal generation priors learned by pretrained T2V models. We therefore initialize the TI2V model parameters W_ϕ from the pretrained T2V parameters W_θ, while introducing and optimizing only a small number of additional parameters for TI2V adaptation. Motivated by this observation, we propose TI2V-Adapter, an efficient adaptation method that can be flexibly applied to diverse pretrained T2V models, enabling their transfer to TI2V generation without full-model retraining.

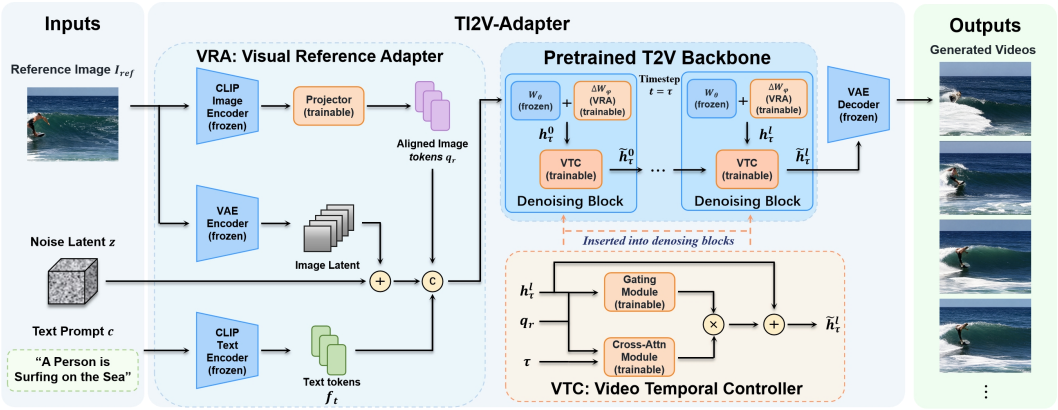

Figure 2: The overall pipeline of the proposed TI2V-Adapter on a frozen T2V backbone.

3.2 Overall Framework

As illustrated in Fig. 2, TI2V-Adapter enables pretrained T2V models to effectively utilize reference-image information through a Visual Reference Adapter (VRA) and a Video Temporal Controller (VTC). Given a frozen T2V backbone, the reference image is first encoded by VRA along two paths: one at the input level and the other at the feature level. At the input level, the reference-image latent is fused with the noise latent to form the initial denoising representation, providing direct visual guidance from the start of generation. At the feature level, the reference image is projected into reference-image tokens, which are further utilized through lightweight trainable adaptation parameters ΔW_ϕ within the frozen T2V backbone. After reference-image information is incorporated into the backbone, VTC further examines the intermediate features to detect weakened reference-image cues and selectively reinforce them throughout the denoising process. During training, only the parameters of TI2V-Adapter are optimized, while the pretrained T2V backbone remains frozen, thereby reducing both optimization cost and computational overhead. It is worth noting that TI2V-Adapter is compatible with existing diffusion-based T2V frameworks and can be readily integrated into diverse pretrained T2V backbones.

3.2.1 Visual Reference Adapter

The proposed Visual Reference Adapter (VRA) aims to parameter-efficiently equip pretrained T2V models with the ability to perceive and utilize reference-image cues. More specifically, VRA incorporates reference-image information at both the input and feature levels, allowing the model to use image cues throughout the video generation process.

At the input level, instead of initializing the diffusion process from pure Gaussian noise as in conventional diffusion-based T2V models, we incorporate reference-image information into the initial latent representation as:

$$z_E = (1-\mu)z + \mu\mathcal{I}(E_{\text{vae}}(I_{ref}), z), \tag{3}$$

where z_E denotes the image-augmented input noise variable, and $E_{\text{vae}}(\cdot)$ denotes a pretrained variational autoencoder (VAE) (Esser et al., 2021), and μ controls the strength of reference-image guidance. Specifically, $\mathcal{I}(\cdot, z)$ resizes and temporally expands the reference-image latent from $\mathbb{R}^{C\times H\times W}$ to $\mathbb{R}^{C\times T\times H\times W}$ to match the shape of z.

Although input-level guidance introduces reference-image information into the generation process, the pretrained T2V model is not designed to effectively utilize such information during denoising. We therefore introduce lightweight feature-level adaptation. Specifically, the reference image is first encoded by a frozen CLIP image encoder (Radford et al., 2021):

$$f_r = \text{CLIP}_{\text{img}}(I_{ref}), \tag{4}$$

where $f_r \in \mathbb{R}^{K\times d}$ denotes the extracted reference-image feature, K and d represent the number of image tokens and the feature dimension, respectively. Following the same encoding process, the

corresponding text feature f_t is extracted using a frozen CLIP text encoder as:

$$f_t = \text{CLIP}_{\text{txt}}(c). \tag{5}$$

To better capture reference-image information and align its feature scale with that of the text features, we employ a lightweight projector to transform f_r into the conditioning space of the pretrained T2V backbone:

$$q_r = s \cdot P_\phi(f_r), \tag{6}$$

where $P_\phi(\cdot)$ denotes the trainable projector, s is a scaling factor, and q_r denotes the projected reference-aware image tokens. The projected image tokens are then concatenated with the text features, *i.e.* $[q_r, f_t]$, and fed into the adaptation layers.

The adaptation layers are designed to jointly consider textual semantics and reference-image cues, enabling the pretrained T2V backbone to generate videos that are both text-aligned and image-consistent. Specifically, each adaptation layer bridges the pretrained weight W_θ in the corresponding layer with a lightweight trainable update ΔW_ϕ:

$$W' = W_\theta + \Delta W_\phi \tag{7}$$

For parameter-efficient adaptation, ΔW_ϕ is parameterized in a low-rank form:

$$\Delta W_\phi = \frac{\alpha}{r} W_\phi^1 W_\phi^2. \tag{8}$$

W_ϕ^1 and W_ϕ^2 denote trainable low-rank matrices that project features into and out of an r-dimensional subspace, respectively. r denotes the rank of the low-rank decomposition of ΔW_ϕ, and α is a scaling factor.

With the proposed VRA augmenting the pretrained T2V model at both the input and feature levels, the image-conditioned generation process can be formulated as

$$V = G(z_E, f_t, q_r; W_\theta + \Delta W_\phi). \tag{9}$$

In this way, VRA introduces parameter-friendly, learnable reference-image conditioning into the pretrained T2V backbone while preserving its original text-video and temporal generation priors.

3.2.2 Video Temporal Controller

Although VRA injects reference-image information into the generation process, reference-image cues may gradually weaken over diffusion timesteps, causing the generated content to progressively deviate from the image appearance, as shown in Fig. 3. To mitigate this issue, we devise the Video Temporal Controller (VTC) to regulate the propagation of reference-image cues along the denoising trajectory, preventing the gradual attenuation of image information during generation.

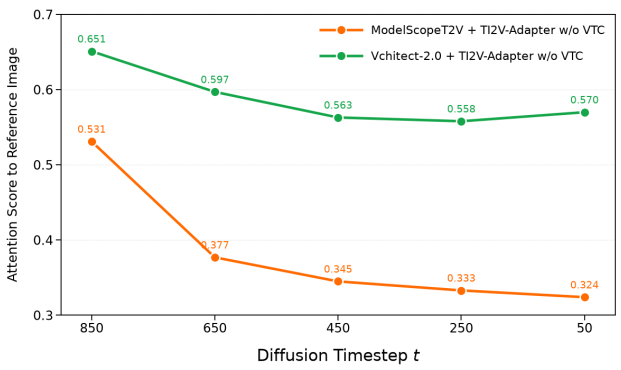

Figure 3: Comparison of reference-image information propagation without the proposed Video Temporal Controller.

Let h_τ^l denote the feature map of the l-th middle block at diffusion step τ, and q_r denote the projected reference-aware image tokens. To determine how much reference-image information should be reinforced at each denoising step, we predict a gating coefficient from the current hidden state h_τ^l, image tokens q_r, and timestep information τ:

$$g_\tau^l = \sigma\left(\text{MLP}\left[\text{Pool}(h_\tau^l), \text{Pool}(q_r), e(\tau)\right]\right), \tag{10}$$

where $\text{Pool}(\cdot)$ denotes the pooling operation, $e(\tau)$ denotes the timestep embedding, and MLP denotes a Multi-Layer Perceptron. Based on the estimated gating coefficient, we further compute

image-aware temporal updates using cross-attention (Vaswani et al., 2017), *i.e.* Attn, between the current denoising state and the image tokens:

$$\Delta h_\tau^l = \mathrm{Attn}\left(h_\tau^l W_q, q_r W_k, q_r W_v\right) W_o, \tag{11}$$

where W_q, W_k, W_v, W_o are learnable parameters. The attention weights measure the relevance between the current denoising state and the image tokens, allowing the model to select image-aware information to compensate for the weakened image cues in h_τ^l.

Finally, the controlled hidden state is updated as:

$$\tilde{h}_\tau^l = h_\tau^l + \lambda_\tau^l \, g_\tau^l \odot \Delta h_\tau^l, \tag{12}$$

where λ_τ^l denotes a predefined scaling coefficient that modulates the injection strength of the image signal across different denoising stages.

With this procedure, VTC mitigates image drift along the denoising trajectory by adaptively reinforcing weakened image cues in intermediate hidden states, adding only a lightweight MLP and cross-attention module as trainable components.

3.3 Training Objective

During training, we freeze the original diffusion-based T2V backbone and update only the parameter-efficient adaptation modules introduced by TI2V-Adapter. We adopt the standard diffusion denoising objective. Given a clean latent representation z_0 of the ground-truth video, the noisy latent variable z_t at timestep t is obtained through the forward diffusion process as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{13}$$

where $\bar{\alpha}_t$ denotes the cumulative noise schedule.

At timestep t, the TI2V model predicts the added noise conditioned on the image-aware conditioning tokens, leading to the following denoising objective:

$$\mathcal{L}_{\mathrm{denoise}} = \mathbb{E}_{z_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_{W_\theta, W_{\theta\to\phi}}(z_t, q_r, f_t, t)\right\|_2^2\right]. \tag{14}$$

Here, $\epsilon \sim \mathcal{N}(0, I)$ denotes Gaussian noise, and $\epsilon_{W_\theta, W_{\theta\to\phi}}$ denotes the noise prediction network built upon the pretrained T2V backbone. We keep the pretrained parameters W_θ frozen, and optimize only $W_{\theta\to\phi}$, which comprises the trainable parameters of VRA and VTC.

4 Experiments

4.1 Datasets and Metrics

4.1.1 Datasets

We quantitatively and qualitatively evaluate TI2V-Adapter on the UCF101 (Soomro et al., 2012) and Kinetics-600 (Carreira et al., 2018) datasets. UCF101 contains 101 human action categories and 13,320 video clips. Following TI2V-Zero (Ni et al., 2024), we select the same nine action categories from UCF101, using 935 videos for training and 1,000 video clips for testing. Similarly, Kinetics-600 covers 600 human action categories. To ensure a fair comparison of data efficiency across datasets, we select 12 action categories and also sample 935 clips for training and 1,000 for testing. For both datasets, the first frame of each video is used as the reference image, and the corresponding action label is inserted into the template *"A person is [action description]."* to construct the text prompt. For example, the label *surfing* is converted to *"A person is surfing."*

4.1.2 Evaluation Metrics

Following existing video generation works, we adopt the Fréchet Video Distance (FVD) to evaluate the overall visual quality, temporal coherence, and distribution diversity of generated videos. FVD (Unterthiner et al., 2019) measures the Fréchet distance between feature distributions of real and generated videos extracted by a pretrained I3D network (Carreira & Zisserman, 2017). Furthermore, we utilize text-conditioned FVD (tFVD) (Ni et al., 2024) to measure video generation quality

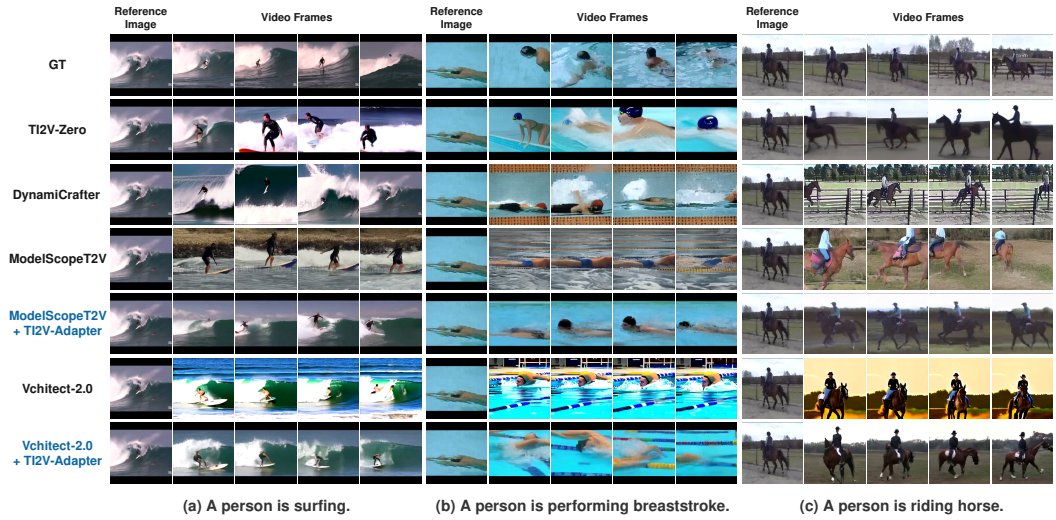

Figure 4: Qualitative comparison of TI2V generation results on the UCF101 dataset. The first column shows the reference image, while the remaining columns show the 2nd, 6th, 11th, and 16th frames of each video. Text conditions are provided below each example.

Table 1: Quantitative comparison on UCF101 and Kinetics-600 for TI2V generation. The training cost is measured by trainable parameters and training samples. **Bold** and underline indicate the best and second-best results among non-training-intensive methods, respectively.

Method	Training Cost		UCF101		Kinetics-600	
	Trainable Params	Samples	FVD↓	tFVD↓	FVD↓	tFVD↓
Pretrained T2V Models						
Vchitect-2.0	-	-	1981.65	3412.37	1615.10	3050.63
ModelScopeT2V	-	-	970.51	2192.87	763.05	1879.59
Existing TI2V Generation Methods						
DynamiCrafter	1.49B	10M	656.71	1689.37	347.42	1363.08
TI2V-Zero	0	0	477.19	1306.75	522.12	1505.61
ConsistI2V	1.2B	10M	348.79	1032.54	436.42	1320.59
Efficient adaptation from T2V to TI2V (Ours)						
Vchitect-2.0+ TI2V-Adapter	6.01M	935	745.41	1695.18	1008.69	2303.10
ModelScopeT2V + TI2V-Adapter	9.90M	935	<u>404.16</u>	<u>1296.64</u>	<u>356.91</u>	<u>1315.35</u>
TI2V-Zero + TI2V-Adapter	9.90M	935	**401.60**	**1286.92**	**343.53**	**1283.96**

under different text conditions. Specifically, tFVD is computed by calculating FVD separately within each text condition and then averaging the scores across all conditions. For both FVD and tFVD, a lower value indicates better performance. All metrics are computed based on the full set of 1,000 test samples.

4.2 Implementation Details

We train and evaluate TI2V-Adapter in PyTorch using four NVIDIA GeForce RTX 5090 GPUs. As a plug-and-play module, TI2V-Adapter optimizes only the newly introduced lightweight adaptation parameters while keeping the pretrained T2V backbone frozen. The model is trained for 35 epochs on both datasets using the AdamW optimizer (Loshchilov & Hutter, 2019). The learning rates are set to 8×10^{-5} and 6×10^{-5} for UCF101 and Kinetics-600, respectively, with zero weight decay and gradient clipping at 0.75. We adopt DDIM sampling (Song et al., 2021a) with classifier-free guidance (Ho & Salimans, 2022) during inference. The rank r of the low-rank adaptation in the Visual Reference Adapter (VRA) is set to 4.

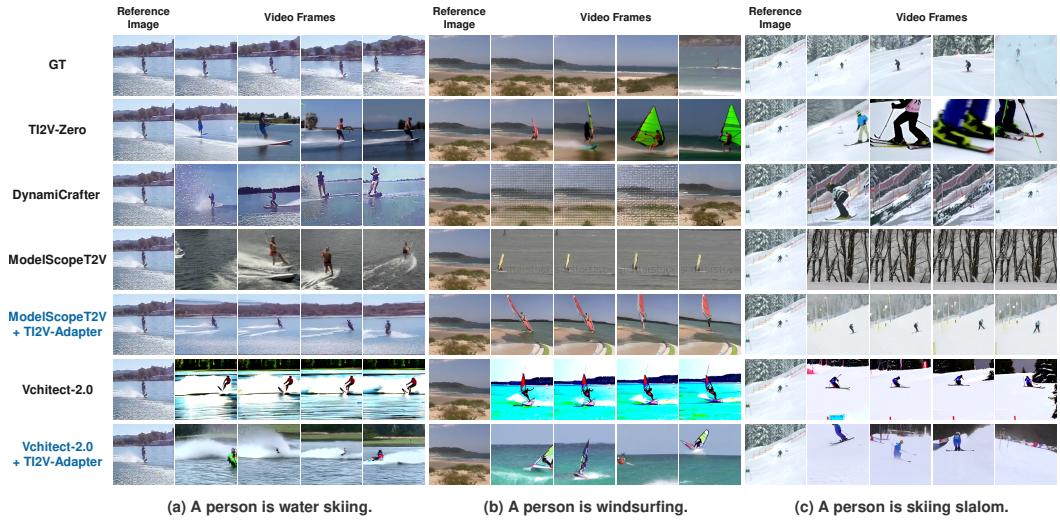

Figure 5: Qualitative comparison of TI2V generation results on Kinetics-600. The first column shows the reference image, while the remaining columns show the 2nd, 6th, 11th, and 16th frames of each video. Text conditions are provided below each example.

4.3 Comparison with the State-of-the-Art Methods

We compare TI2V-Adapter with representative methods from three categories. These include pretrained T2V models such as ModelScopeT2V (Wang et al., 2023) and Vchitect-2.0 (Fan et al., 2025), the training-free TI2V method TI2V-Zero (Ni et al., 2024), and training-intensive TI2V methods including DynamiCrafter (Xing et al., 2024b) and ConsistI2V (Ren et al., 2024). As TI2V-Adapter is architecture-agnostic, we further evaluate its generalizability by integrating it with different pretrained T2V models. The quantitative results are shown in Table 1. The pretrained T2V models achieve poor performance in the TI2V setting, mainly because they are not explicitly designed to perceive and utilize reference-image conditions. After integrating TI2V-Adapter, the FVD/tFVD scores of Vchitect-2.0 decrease from 1981.65/3412.37 to 745.41/1695.18, while those of ModelScopeT2V decrease from 970.51/2192.87 to 404.16/1296.64, respectively. Furthermore, TI2V-Adapter can be seamlessly integrated with TI2V-Zero, consistently improving its video generation performance.

Even compared with training-intensive methods, TI2V-Adapter remains highly competitive while requiring substantially fewer trainable parameters and training samples. Specifically, ModelScopeT2V + TI2V-Adapter achieves an FVD of 404.16 on UCF101 with only 9.90M trainable parameters and 935 training samples, clearly outperforming DynamiCrafter (656.71 FVD, 1.49B parameters, 10M samples). The qualitative results in Fig. 4 and Fig. 5 further demonstrate that TI2V-Adapter improves reference-image consistency and temporal coherence. These consistent improvements across different pretrained models indicate that TI2V-Adapter does not rely on a specific backbone architecture, but provides a general adaptation mechanism for enhancing reference-image conditioning.

4.4 Ablation Studies

For a fair comparison, all ablation studies are conducted on the UCF101 dataset using ModelScopeT2V as the backbone, while keeping the training configurations consistent with those described in the implementation details.

Ablation Study 1: Effectiveness of Core Components. As shown in Tab. 2, we conduct ablation studies to evaluate the effectiveness of VRA and VTC in the proposed TI2V-Adapter. Incorporating input-level reference-image guidance improves the generation performance, while further introducing feature-level guidance brings additional gains, reducing the FVD and tFVD to 481.69 and 1480.55, respectively. Moreover, introducing VTC can further enhance the video generation performance to 404.16 (FVD) and 1296.64 (tFVD). These progressive improvements validate the complementary roles of the two-level reference-image adaptation in VRA and the VTC.

Table 2: Ablation study on the effectiveness of the VRA and VTC modules in TI2V-Adapter.

Setting	VRA		VTC	FVD↓	tFVD↓
	Input	Feature			
A	–	–	–	970.51	2192.87
B	✓	–	–	693.84	1663.26
C	✓	✓	–	481.69	1480.55
D	✓	✓	✓	**404.16**	**1296.64**

Table 3: Ablation study on different encoder pairings on UCF101.

Encoder Configuration		UCF101	
Feature-level	Input-level	FVD↓	tFVD↓
CLIP	CLIP	767.78	1810.10
VAE	VAE	422.72	1346.23
VAE	CLIP	803.50	1865.83
CLIP	VAE	**404.16**	**1296.64**

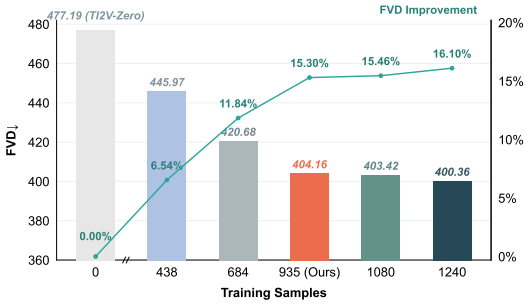

Figure 6: Effect of training data size on UCF101.

Table 4: Effect of the rank r on UCF101.

Rank r	Params (M)	FVD↓	tFVD↓
2	5.0M	417.58	1309.66
4	9.90M	404.16	1296.64
8	15.4M	429.65	1327.25
16	29.3M	441.73	1341.42

Ablation Study 2: Effect of Encoder Configurations. As shown in Fig. 2, CLIP and VAE provide feature-level and input-level reference-image cues, respectively. We further evaluate different encoder configurations in Tab. 3. Using CLIP for extracting feature-level image cues and VAE for input-level guidance achieves the best performance. This may be because CLIP and VAE provide complementary visual information: CLIP is more suitable for feature-level conditioning due to its high-level semantic representations, while VAE preserves low-level appearance and structural information that is better suited for input-level latent augmentation.

Ablation Study 3: Effect of the Number of Training Samples. As shown in Fig. 6, we evaluate TI2V-Adapter with different numbers of training samples to investigate its data efficiency. To construct larger training sets beyond the original 935 samples, we additionally use videos from the validation/evaluation pools that do not overlap with the testing set. With only 935 training samples, TI2V-Adapter achieves a **15.3%** FVD improvement over the training-free TI2V-Zero baseline, while further increasing the training data yields only marginal gains. Therefore, we use 935 training samples as the default setting for efficient TI2V adaptation.

Ablation Study 4: Effect of the rank r. As shown in Tab. 4, we evaluate the effect of different rank settings on the performance and parameter efficiency of TI2V-Adapter. $r = 4$ achieves the best performance. We attribute this result to the fact that a smaller r provides insufficient adaptation capacity to effectively incorporate reference-image information, while a larger r introduces redundant parameters without further improving the adaptation capability. Therefore, we adopt $r = 4$ as the default setting, achieving the best generation performance with only 9.90M trainable parameters.

5 Conclusion

This paper proposes TI2V-Adapter, a plug-and-play and parameter-efficient adaptation method for transferring pretrained T2V models to TI2V generation models. TI2V-Adapter can be efficiently integrated into various T2V backbones, preserving their original text-driven generation capability while effectively extending them to image-conditioned video generation with only limited training data and trainable parameters. Extensive experiments demonstrate the superior performance and efficiency of TI2V-Adapter on UCF101 and Kinetics-600 datasets. In future work, we will extend TI2V-Adapter to long-video generation, where maintaining reference-image consistency over extended temporal horizons remains particularly challenging.

REFERENCES

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 (CVPR)*, pp. 22563–22575, 2023.

João 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 (CVPR)*, pp. 6299–6308, 2017.

João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8VG8tpPZhe.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883, 2021.

Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025.

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. In *International Conference on Learning Representations (ICLR)*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 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.

Chenyu Hui, Xiaodi Huang, Siyu Xu, Yunke Wang, Shan You, Fei Wang, Tao Huang, and Chang Xu. Seeing realism from simulation: Efficient video transfer for vision-language-action data augmentation. In *International Conference on Machine Learning*, 2026.

Bolin Lai, Sangmin Lee, Xu Cao, Xiang Li, and James M. Rehg. Unified text-image-to-video generation: A training-free approach to flexible visual conditioning. *arXiv preprint arXiv:2505.20629*, 2025.

Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on Artificial Intelligence (AAAI)*, volume 32, 2018.

Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, et al. Stiv: Scalable text and image conditioned video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

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 (CVPR)*, pp. 9015–9025, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, volume 139, pp. 8162–8171. PMLR, 2021.

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 *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *Transactions on Machine Learning Research (TMLR)*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

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 *International Conference on Learning Representations (ICLR)*, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021b.

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.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. In *International Conference on Learning Representations (ICLR) Workshop*, 2019.

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

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.

Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncrafter: Generative cartoon interpolation. *arXiv preprint arXiv:2405.17933*, 2024a.

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 (ECCV)*, pp. 399–417. Springer, 2024b.

Jinfan Zhou, Lixin Luo, Sungmin Eum, Heesung Kwon, and Jeong Joon Park. Generative spatiotemporal data augmentation. *arXiv preprint arXiv:2512.12508*, 2025.

AI Assistance Disclosure

In this work, we used generative AI tools to improve the readability and grammar of the manuscript, assist with brainstorming and organizing the presentation of the paper, and support literature retrieval and discovery of related work. We did not use generative AI tools to design the methodology, formulate research hypotheses, conduct experiments, implement the proposed method, generate or interpret experimental results, or produce the scientific figures. All AI-assisted content was reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, citations, and artifacts.