Prompt Learning Based Adaptor for Enhanced Video Editing with Pretrained Text-to-Image Diffusion Models

Yangfan He*
University of Minnesota - Twin Cities
he000577@umn.edu

Sida Li
Peking University
lisida0128@gmail.com

Jianhui Wang
University of Electronic Science and Technology of China
2022091605023@std.uestc.edu.cn

Abstract

The rapid advancement of text-to-image generation technologies based on diffusion models has produced remarkable results, driving the emergence of video editing applications built upon these pretrained models. To achieve temporal consistency with independent framewise text-to-image generation, existing video editing models either fine-tune temporal layers or propagate temporal features at test time without additional training. While these approaches show promise, the frame independence of text-to-image generation creates a bottleneck in delivering consistent and high-quality video outputs.In this paper, we propose a lightweight adaptor utilizing prompt learning to enhance video editing performance with minimal training cost. Our approach introduces shared prompt tokens to improve editing capabilities and unshared frame-specific tokens to impose consistency constraints across frames. The adaptor seamlessly integrates into existing video editing pipelines, offering significant improvements in temporal coherence and overall video quality, benefiting a broad spectrum of downstream video editing algorithms.

1 Introduction

With the rapid advancement of text-to-image (T2I) diffusion models, text-to-video (T2V) models have also made significant strides. However, existing T2V models face challenges such as high training costs, complex algorithms, and large parameter sets, which limit their practical application. In this study, we propose a prompt-learning-based adapter to fine-tune T2I models, addressing these challenges specifically for video editing tasks. Our approach integrates two key strategies to enhance video quality and temporal consistency. First, we introduce a lightweight adapter that leverages prompt learning to fine-tune pretrained T2I models, utilizing shared and unshared token mechanisms to efficiently adapt to various video editing requirements. Second, we incorporate a cosine similarity constraint between adjacent frames during early denoising steps, in addition to the mean squared error (MSE) loss on noise. This prompt-learning-based adapter, with the loss constraints of consistency of adjacent frame in early diffusion denoising period, enhances video quality and temporal coherence. Experimental results demonstrate that our method not only reduces training costs but also achieves notable improvements in temporal consistency, image quality, and text alignment in T2V generation, offering an efficient and practical solution for video editing.

*Corresponding author.

The NeurIPS 2024 Workshop on Adaptive Foundation Models (AFM 2024), Vancouver, Canada, 2024.

- We propose a novel prompt-learning based adapter that is plug-and-play, trained on T2I models using LoRA for fine-tuning. This approach is highly efficient and low-cost, making it adaptable for various downstream applications.
- To improve temporal coherence when training, we introduce a consistency constraint between adjacent frames during early denoising, leveraging temporal depth information to enhance video quality and frame continuity. Furthermore, we propose a shared token strategy that integrates shared tokens with text embeddings and utilizes unshared tokens for low-resolution frame constraints, thereby improving consistency throughout the video generation process.
- Experimental results indicate that the proposed adapter enhances the performance of downstream video editing algorithms, delivering improved text-to-image alignment, high image fidelity, and consistent frame coherence.

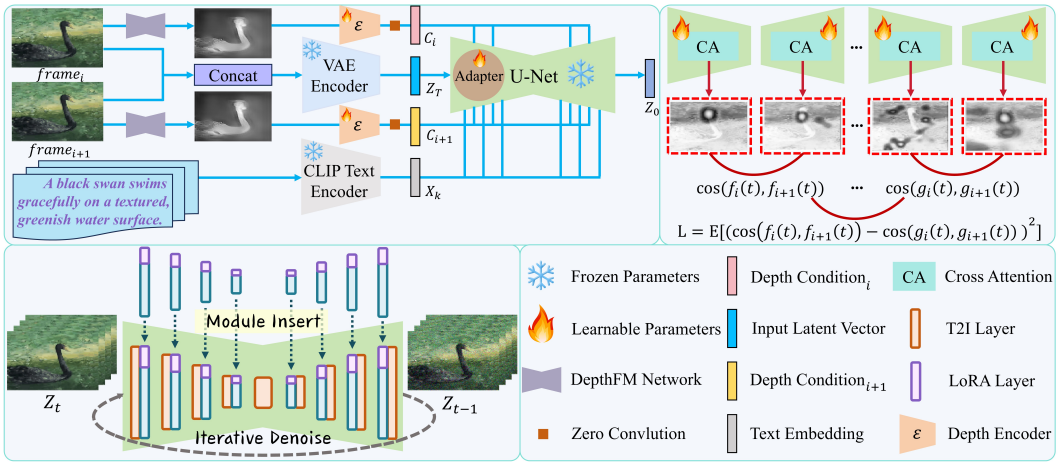

Figure 1: An overview of the video generation process using LoRA-enhanced feature extraction. Depth and text embeddings are combined with latent vectors and processed through iterative denoising, cross attention, and cosine similarity constraints between adjacent frames.

2 Related Work

Text-to-image (T2I) diffusion models have driven advancements in text-to-video (T2V) generation, classified into three paradigms: Inversion and Sampling, Upstream, and Downstream. The Inversion and Sampling paradigm improves efficiency by modifying diffusion mechanisms, using techniques like latent diffusion models [7] and ODE solvers [9]. The Upstream approach incorporates adapters into T2I models, facilitating fine-tuning and enhancing video editing [13]. The Downstream approach refines architectures like Unet [8] and ControlNet [13], employing advanced attention mechanisms for coherence and video quality. Methods like Gen-L-Video [10], FLATTEN [2], and StableVideo [1] address T2I-based video editing but face challenges in validation, structural changes, and non-rigid deformations. Low-Rank Adaptation (LoRA) techniques [4] further improve T2I models by enabling personalized fine-tuning through block-wise and dynamic feature adjustments. These developments underscore the evolving landscape of video generation and editing.

3 Method

Achieving temporal and semantic coherence in video generation requires effective fine-tuning. We enhance the U-Net architecture by integrating Low-Rank Adaptation (LoRA) and registering hooks to extract intermediate features at each forward pass. For each U-Net block (l, b) at time step t, feature extraction is modified as:

$$\mathbf{F}_{l,b}^{t} = \begin{cases} \mathbf{W}_0\mathbf{x}, & \text{if rank}(\mathbf{A}_{l,b}) = 0, \\ \mathbf{W}_0\mathbf{x} + \mathbf{B}_{l,b}\mathbf{A}_{l,b}\mathbf{x}, & \text{otherwise}, \end{cases} \tag{1}$$

where $\mathbf{x}$ is the input feature vector. The projection matrices for query ($\mathbf{W}_{q,l,t}$), key ($\mathbf{W}_{k,l,t}$), and value ($\mathbf{W}_{v,l,t}$) are updated accordingly. Temporal coherence is enforced using similarity constraints on adjacent frames:

$$\text{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1}) = \frac{\mathbf{F}_t \cdot \mathbf{F}_{t+1}}{\|\mathbf{F}_t\| \|\mathbf{F}_{t+1}\|} \tag{2}$$

The temporal loss $\mathcal{L}_{\text{temporal}}$ is formulated to minimize the difference in similarity across consecutive frames:

$$\mathcal{L}_{\text{temporal}} = \frac{1}{T} \sum_{t=1}^{T-1} \left(\text{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1}) - \text{Sim}(\mathbf{F}_{t-1}, \mathbf{F}_t)\right)^2 \tag{3}$$

This constraint keeps features aligned between consecutive frames, reducing flickering and improving content consistency across the video.

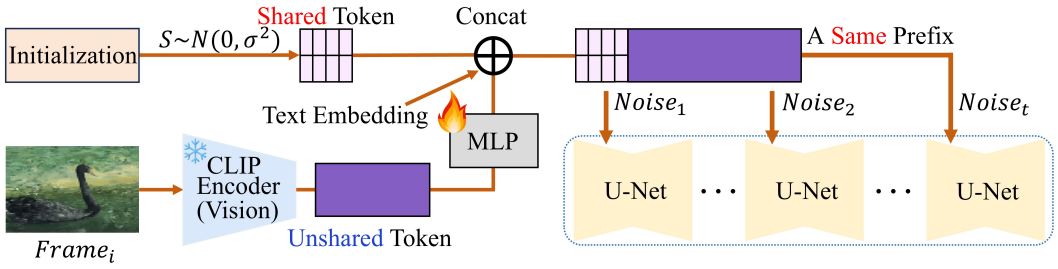

Figure 2: Shared and unshared token mechanism for video generation. Shared tokens ensure global consistency across frames, while unshared tokens handle frame-specific details. The same prefix is applied across all time steps, and only shared tokens are updated during the final phase.

To enhance text-to-video semantic alignment, we train shared tokens for multi-frame consistency to improve image-text alignment, combining them with conditional and unconditional embeddings to create unified text embedding guidance. The shared tokens are initialized from a normal distribution:

$$T_{\text{share}} \sim \mathcal{N}(0, 0.02), \quad T_{\text{share}} \in \mathbb{R}^{N_{\text{share}} \times 768} \tag{4}$$

Given an input image I, the visual hidden representation is extracted using the CLIP model's vision encoder:

$$H_{\text{vision}} = f_{\text{clip\_vision}}(I) \tag{5}$$

Subsampling is then applied to downsample the hidden states to obtain the unshared tokens:

$$H_{\text{sub}} = \begin{cases} H_{\text{vision}} & \text{if } s = 1, \\ \text{Concat}\left(H_{\text{vision}}[:, :1, :], H_{\text{vision}}[:, 1::s, :]\right) & \text{if } s > 1, \end{cases} \tag{6}$$

The embedding tokens are subsampled along the sequence dimension, retaining the first token to capture key information. The resulting visual representation, H_{sub}, is projected through a linear layer to align the unshared tokens in the same latent space as the token embeddings previously processed by the CLIP model:

$$Z_{\text{unshare}} = W_{\text{proj}} \cdot H_{\text{sub}} \tag{7}$$

The unshared token incorporates low-resolution information to enhance frame quality, while the shared token ensures consistency across adjacent frames, and all tokens—shared, unshared, conditional, and unconditional—are concatenated for LoRA fine-tuning:

$$Z_{\text{final}} = \text{Concat}([T_{\text{share}}, Z_{\text{unshare}}, Z_{\text{cond}}, Z_{\text{uncond}}], \text{dim} = 0) \tag{8}$$

During training, the CLIP vision encoder (θ) remains frozen while the LoRA parameters, shared embeddings, and non-shared token projection layers (ϕ) are updated iteratively to minimize the loss:

$$\text{Loss} = \begin{cases} \left|\epsilon - \epsilon_\theta(x_t, t, D, Z_{\text{final}})\right|^2, & \text{if } t \in [0, 0.9T], \\ \left|\epsilon - \epsilon_\theta(x_t, t, D, Z_{\text{final}})\right|^2 + \lambda \mathcal{L}_{\text{lora}}, & \text{if } t \in [0.9T, T], \end{cases} \tag{9}$$

4 Experiments

4.1 Quantitive Results

As shown in Table 1, the comparison between video generation algorithms with and without LoRA highlights improvements in perceptual quality and text-image alignment. Lower LPIPS and FID values indicate better visual fidelity, while higher CLIP scores suggest improved alignment with text prompts. The results demonstrate that the inclusion of LoRA generally leads to better performance, with the best outcomes highlighted in gray, showing enhanced image quality and semantic coherence across the tested algorithms.

Table 1: Comparison of Algorithms with and without LoRA for LPIPS, CLIP, and FID metrics on the MSR-VTT [12] testing set. Adapter parameters were injected during the 0.9 - 1.0 time intervals of denoising. Best-performing results are highlighted with gray.

Algorithm	**LoRA**	**LPIPS↓**	**CLIP↑**	**FID↓**
Text2Video-Zero [5]	w	0.319	31.38	–
	w/o	0.402	28.82	–
TokenFlow [3]	w	0.129	29.15	143.21
	w/o	0.135	28.90	166.72
Vid2Vid [11]	w	0.253	32.78	154.3
	w/o	0.303	31.94	149.4
VidToMe [6]	w	0.122	29.94	143.33
	w/o	0.123	29.93	146.27

Table 2 summarizes the impact of removing depth maps, shared tokens, and LoRA fine-tuning on the performance of different video generation algorithms. The results indicate that the combination of all components (depth maps, shared tokens, and LoRA) consistently yields the best performance across FID, CLIP, and LPIPS metrics, highlighting the significance of each component in improving video quality and semantic alignment.

Table 2: Ablation Study: Effects of Depth Maps, Shared Tokens, and LoRA Fine-Tuning on Downstream Algorithms

Configuration	**Algorithm**	**FID↓**	**CLIP↑**	**LPIPS↓**
w/o Depth Maps	Tokenflow [3]	164.60	28.23	0.131
	Vid2Vid [11]	202.14	32.06	0.288
	VidToMe [6]	159.20	29.74	0.134
	Text2Video [5]	–	29.37	0.381
w/o Shared Tokens	Tokenflow [3]	190.50	27.30	0.138
	Vid2Vid [11]	231.05	32.80	0.261
	VidToMe [6]	168.00	29.50	0.121
	Text2Video [5]	–	28.10	0.397
All Components (Base)	Tokenflow [3]	143.21	29.15	0.129
	Vid2Vid [11]	154.37	32.78	0.253
	VidToMe [6]	143.33	29.89	0.121
	Text2Video [5]	–	31.38	0.380

4.2 User Study

We conducted a user study with 67 random participants, independent of age and gender, to evaluate video generation models trained on the MSR-VTT dataset. Each participant was shown video clips generated with and without LoRA fine-tuning for four algorithms. Participants were asked to choose the better video based on three criteria: (1) text-to-image alignment, (2) image quality, and (3) consistency. As shown in Figure 3, LoRA fine-tuning was preferred across most algorithms, particularly for VidToMe in terms of consistency, and TokenFlow for image quality. The results

indicate that LoRA consistently improves both semantic alignment and frame quality, especially for TokenFlow and Vid2Vid, where the improvements are more noticeable.

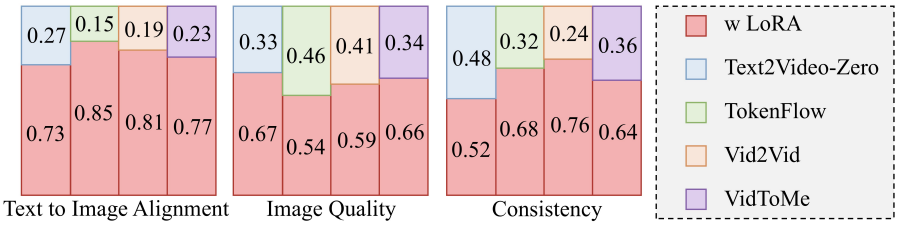

Figure 3: Comparison of user study results between video generation algorithms (Text2Video-Zero, TokenFlow, Vid2Vid, and VidToMe) with and without LoRA fine-tuning. The results show the percentage of user preferences for each method across the different algorithms.

4.3 Visualization Results

Figure 4 presents a qualitative comparison of video frames generated by four different algorithms (Text2Video-Zero, TokenFlow, Vid2Vid, and VidToMe) with and without Low-Rank Adaptation (LoRA). The results demonstrate that incorporating LoRA improves both the coherence and quality of the generated videos across all algorithms.

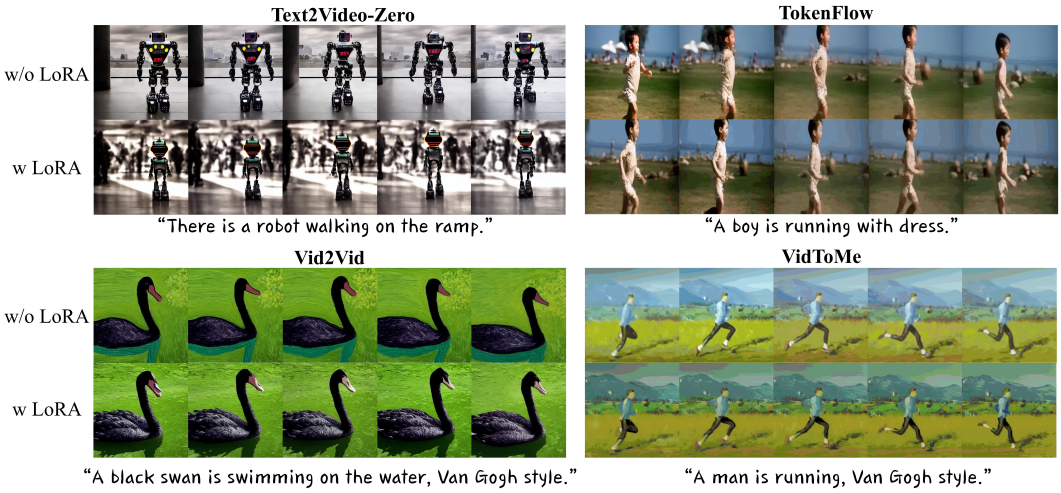

Figure 4: Comparison of video generation quality between different algorithms (Text2Video-Zero, TokenFlow, Vid2Vid, and VidToMe) with and without LoRA fine-tuning.

5 Conclusion and Future Work

In this paper, we introduced a prompt-learning based adaptor for video editing using pretrained text-to-image diffusion models, leveraging Low-Rank Adaptation (LoRA) and a shared/unshared token mechanism to significantly enhance temporal consistency and video quality with minimal training costs. Our plug-and-play solution effectively integrates with existing models, improving text-image alignment and reducing flickering. Future work will focus on adaptive LoRA injection strategies, testing the approach across a broader range of video editing algorithms (both training-free and those requiring training), extending to higher-resolution video generation, designing a more generalized personalized adapter, and further improving both performance metrics and visual quality.

References

[1] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

[2] Y. Cong, M. Xu, christian simon, S. Chen, J. Ren, Y. Xie, J.-M. Perez-Rua, B. Rosenhahn, T. Xiang, and S. He. FLATTEN: optical FLow-guided ATTENtion for consistent text-to-video editing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JgqftqZQZ7.

[3] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel. Tokenflow: Consistent diffusion features for consistent video editing, 2023. URL https://arxiv.org/abs/2307.10373.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

[5] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators, 2023. URL https://arxiv.org/abs/2303.13439.

[6] X. Li, C. Ma, X. Yang, and M.-H. Yang. Vidtome: Video token merging for zero-shot video editing, 2023. URL https://arxiv.org/abs/2312.10656.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. URL https://arxiv.org/abs/2112.10752.

[8] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

[9] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

[10] F.-Y. Wang, W. Chen, G. Song, H.-J. Ye, Y. Liu, and H. Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising, 2023. URL https://arxiv.org/abs/2305.18264.

[11] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis, 2018. URL https://arxiv.org/abs/1808.06601.

[12] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. doi: 10.1109/CVPR.2016.571.

[13] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

A Appendix

A.1 Training Settins

Training adapter is mainly based on Stable Diffusion v1.5, which contains 860M UNet and 123M text encoder, suitable for GPUs with at least 10GB VRAM. We first trained a ControlNet based on Stable Diffusion v1.5 using the original image input and depth map to provide depth information for subsequent UNet training. At the same time, we trained a PrefixToken module separately to concatenate the shared token with the features extracted from the image, which helps to align the generated video and the input prompt in the semantic space. Finally, we fixed the parameters of the UNet and text encoder and added a trainable adapter to the UNet downsampling block. During training, we only select timesteps in the range of 0.9 to 1.0 for DDPM, and use the pre-trained

sharetoken to replace the original text input with probability at all times, with the probability set to 0.3. The learning rate is controlled at 3e-5, the rank is 4, the mixed precision training of fp16 is adopted, the input image resolution is 512x512, and the loss ratio defaults to 0.01. Noting that the low-rank structure significantly reduces computational overhead, our 10-epoch training on the MSRVTT dataset can be completed in just 20 hours on a single NVIDIA GeForce RTX 4090.

A.2 Inference Settings

The versatility of this method means that when it is applied to downstream tasks, it only needs to load lora weight in StableDiffusionPipeline. We default that the LoRA weight will be integrated into the original model weight at a ratio of 50%, and the LoRA parameters will only be injected when the timesteps are in the range of 0.9-1.0.

A.3 Datasets

Training sets MSR-VTT is a large-scale open-domain video captioning dataset with 10,000 video clips from 20 categories, each annotated with 20 English sentences, totaling approximately 29,000 unique words, and split into 6,513 training clips, 497 validation clips, and 2,990 testing clips. We used the MSR-VTT training set and the OpenAI ChatGPT API to modify video captions and generate new content, employing a specialized DataLoader to concatenate early denoised adjacent frames into single batches, and converting MP4 videos to WebDataset format to optimize training performance.