Enhancing Low-Cost Video Editing with Lightweight Adaptors and Temporal-Aware Inversion

Yangfan He[1], Sida Li[2]*, Jianhui Wang[3], Kun Li[4], Xinyuan Song[5],
Xinhang Yuan[6], Kuan Lu[7], Menghao Huo[8], Jiaqi Chen[9], Keqin Li[9],
Miao Zhang[10], Xueqian Wang[10]†
[1]University of Minnesota—Twin Cities, [2]Peking University, [3]University of Electronic Science and Technology of China, [4]Xiamen University,
[5]Emory University, [6]Washington University in St. Louis, [7]Cornell University, [8]Santa Clara University,
[9]Independent Researcher, [10]Tsinghua University
wang.xq@sz.tsinghua.edu.cn

Recent advancements in text-to-image (T2I) generation using diffusion models have enabled cost-effective video-editing applications by leveraging pre-trained models, eliminating the need for resource-intensive training. However, the frame independence of T2I generation often results in poor temporal consistency. Existing methods address this issue through temporal layer fine-tuning or inference-based temporal propagation, but these approaches suffer from high training costs or limited temporal coherence. To address these challenges, we propose a General and Efficient Adapter (GE-Adapter) that integrates temporal, spatial and semantic consistency with Baliteral Denoising Diffusion Implicit Models (DDIM) inversion. This framework introduces three key components: (1) Frame-based Temporal Consistency Blocks (FTC Blocks) to capture frame-specific features and enforce smooth inter-frame transitions using temporally aware loss functions; (2) Channel-dependent Spatial Consistency Blocks (SCD Blocks) employing bilateral filters to enhance spatial coherence by reducing noise and artifacts; (3) a Token-based Semantic Consistency Module (TSC Module) to maintain semantic alignment through a combination of shared prompt tokens and frame-specific tokens. Extensive experiments on multiple datasets demonstrate that our method significantly improves perceptual quality, text-image relevance, and temporal coherence. The proposed approach offers a practical and efficient solution for text-to-video (T2V) editing. Our code is available in the supplementary materials.

1. Introduction

Recent advances in deep learning have demonstrated remarkable progress across multiple domains, particularly in natural language processing, time series analysis, and computer vision applications [1–7, 7–14]. Building upon these foundations, text-to-video (T2V) generation and editing techniques [15–19] have emerged as a cutting-edge research direction, focusing on synthesizing and manipulating dynamic video content that accurately reflects textual descriptions while maintaining visual coherence, spatial consistency, and temporal continuity. Recent advancements in diffusion models have significantly accelerated progress in T2V tasks, enabling more efficient and effective video generation and editing. Studies on parameter-efficient fine-tuning have demonstrated the potential of adapter-based methods [20–24]

Traditional T2V methods rely on large-scale video datasets to learn spatial and temporal dynamics across diverse scenes. Pioneering works such as VideoGPT [25], CogVideo [26], and Recipe [27] demonstrate the advantages of leveraging extensive datasets to capture fine-grained motion, object interactions, and scene transitions. However, these methods face significant challenges, including high computational costs, dataset biases, and the need for high-quality annotations.

*Equal contribution
†Corresponding Author

Second Conference on Parsimony and Learning (CPAL 2026).

To address these limitations, early approaches leveraging pre-trained text-to-image (T2I) diffusion models [28–31] have emerged as a promising alternative due to their lower computational requirements. Notable examples include Animatediff [32], Tune-A-Video [33], and TokenFlow [34], which offer cost-effective solutions while maintaining acceptable performance. These approaches can be broadly categorized into training-based and training-free strategies.

Figure 1: Visual comparison of generated video frames with and without our adapter applied to different algorithms.

Training-based strategies focus on fine-tuning temporal or attention layers to enhance temporal consistency and improve editing performance. While effective, this approach incurs high training costs and limited scalability. To mitigate these issues, adapter-based methods [35–38] have been introduced. These methods integrate temporal or attention layers to reduce training overhead and improve generalization. However, their relatively large parameter sizes and limited adaptability highlight the need for further optimization. In contrast, training-free strategies, such as TokenFlow [39], propagate features between adjacent frames during inference. This eliminates training costs but can lead to lower-quality outputs.

To achieve a balance between computational efficiency and high-quality video generation, we propose a novel Consistency-Adapter Framework. This framework integrates temporal-spatial and semantic consistency while minimizing training expenses. A hierarchical temporal-spatial coherence module (HTC Module) enhances video quality through two key blocks: (1) Frame Similarity-based Temporal Consistency Blocks (FTC Blocks), which capture frame-specific information and reduce abrupt feature changes using temporally aware loss functions; and (2) Channel-Dependent Spatial Consistency Blocks (SCD Blocks), which use bilateral filtering to decrease noise and artifacts. Additionally, a Token-based Semantic Consistency Module (TSC Module) ensures semantic alignment by using shared prompt tokens for flexible editing and frame-specific tokens to maintain inter-frame consistency.

The main contributions of this paper are as follows:

- We propose a lightweight, plug-and-play consistency-adapter framework that balances computational efficiency and video quality by integrating temporal, spatial, and semantic consistency.
- We introduce a hierarchical temporal-spatial coherence module (HTC Module) featuring Frame Similarity-based Temporal Consistency Blocks (FTC Blocks) and Channel-Dependent Spatial Consistency Blocks (SCD Blocks), ensuring temporal and spatial coherence while reducing noise

and artifacts. We also develop a token-based semantic consistency module (TSC Module) that employs both shared and frame-specific tokens to maintain semantic alignment across frames.

- We demonstrate that our lightweight adapter, with only 0.755M trainable parameters for the UNet portion and 15.4M for the prompt portion (total size: 860M), achieves over 50% efficiency improvement compared to leading T2V models while enhancing temporal consistency, semantic alignment, and video quality.

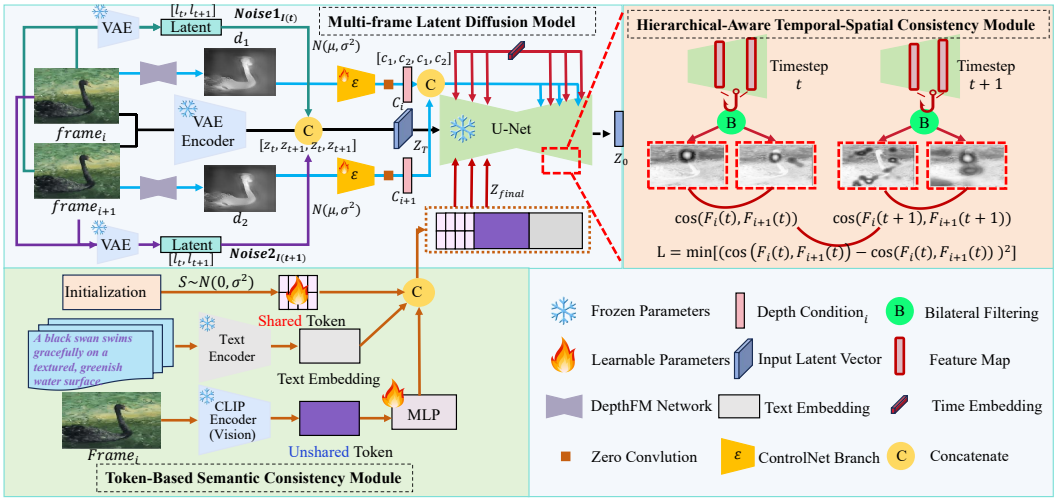

Figure 2: The proposed framework integrates the Multi-frame Latent Diffusion Model, Hierarchical-Aware Temporal-Spatial Consistency Module, and Token-Based Semantic Consistency Module. The top-left illustrates Variational Autoencoder (VAE) encoding, Gaussian noise injection (noise_1, noise_2), and latent concatenation for UNet processing. The top-right shows the Temporal-Spatial Consistency Module, which optimizes frame transitions through a temporal loss. The bottom highlights the Semantic Consistency Module, which combines shared and unshared tokens to improve semantic alignment and reduce flickering, thereby enhancing both spatial and temporal coherence in video generation.

2. Related Work

Text-to-Video Editing. Text-to-Image (T2I) technology has achieved significant advancements through methods like Generative Adversarial Networks (GANs) [40–44] and diffusion models [45–50]. However, extending these advancements to Text-to-Video (T2V) remains challenging. Current T2V approaches include inversion and sampling methods, which optimize diffusion processes [51–54]; upstream methods, which simplify fine-tuning with adapters [55]; and downstream methods, which enhance temporal and semantic consistency using complex attention mechanisms.

For example, methods like Gen-L-Video [56], FLATTEN [57], and StableVideo [58] focus on generating long videos with improved temporal coherence, while approaches like ControlVideo [59] and MagicProp [60] aim to enhance the quality of individual video frames. Despite these advancements, existing 2D U-Net-based T2V models often require training from scratch, freezing pre-trained T2I models and relying heavily on complex temporal layers, leading to computational inefficiency and temporal inconsistency. These methods lack a unified, efficient, and lightweight training paradigm that offers strong generalization with a single run and plug-and-play usability.

DDIM Inversion for Enhanced Video Editing. Denoising Diffusion Implicit Models (DDIM) Inversion [61] exploits the reversibility of DDIMs to control latent space content without regenerating the entire image. This technique enables precise adjustments to object shapes, styles, and details while maintaining consistency. Enhancements like EasyInv [62] and ReNoise [63] refine inversion by iteratively adding and denoising noise, while Eta Inversion [64] introduces a time- and region-dependent η function to improve editing diversity. MasaCtrl [65] identifies object layouts by converting images into noise representations, and Portrait Diffusion [66] merges Q, K, and V values for effective image

blending. While these techniques excel in image editing, there remains a lack of optimized DDIM inversion methods tailored for video generation.

Adapters for Video Editing. Initially developed for natural language processing (NLP) [67], adapters were introduced to efficiently fine-tune large pre-trained models, as demonstrated in BERT [68] and GPT [69]. In computer vision, adapters like ViT-Adapter [37] enable Vision Transformers (ViT) to handle diverse tasks with minimal fine-tuning. Similarly, ControlNet [70] and T2I-Adapter [35] incorporate lightweight modules for diffusion models to provide additional control, while Uni-ControlNet [71] reduces fine-tuning costs with multi-scale conditional adapters for localized control. However, current adapters remain limited to specific architectures and lack the flexibility and generalization required for robust video generation tasks.

3. Method

Text-to-video (T2V) editing requires preserving frame-to-frame temporal coherence, consistent semantics, and spatial structures. Given an input video and text prompts, the goal is to produce an edited video that: (1) **Frame Temporal Alignment**: Video editing should ensure temporal coherence among frames by creating smooth transitions between frames and accurately representing motion dynamics as described by the text prompt; (2) **Video Spatial Alignment**: The edited video should ensure that spatial structures and visual content remain consistent across all frames, aligned with the spatial details described in the text prompt while preserving the integrity of the original videoâĂŹs regions; and (3) **Text-to-video Semantic Alignment**: The algorithm must also ensure that the edited video conveys the intended semantics described in the text prompt while maintaining the contextual consistency of the original video.

We propose a lightweight video adapter that promotes temporal and spatial consistency, as well as semantic alignment, while reducing training costs in 2D UNet-based T2V generation. Our approach integrates:

- A multi-frame latent diffusion model (Section 3.1);
- Frame Similarity-based Temporal Consistency Blocks (FTC Blocks, Section 3.2) to enforce temporal coherence across adjacent frames;
- Channel-dependent Spatial Consistency Blocks (SCD Blocks, Section 3.2) to stabilize noisy latents and reduce frame-level artifacts using bilateral filters;
- A Token-based Semantic Consistency Module (TSC Module, Section 3.3) to align text and video latents effectively.

3.1. Base Diffusion model

This model differs from traditional methods by directly modeling latent features for consecutive video frames [17, 34, 72]. In-depth analyses on the inner workings of diffusion models have also been reported [23], which support our design choices. By incorporating ControlNet branches [70], time embeddings [73], and concatenated latent representations, the model aligns temporal features effectively, reducing flickering and structural inconsistencies. Additionally, the UNetâĂŹs decoder layers are optimized for high-resolution outputs, enabling fine-grained reconstruction of temporal dynamics and appearance details. Overall, the Multi-frame Latent Diffusion Model s serves as the backbone of our video editing framework, generating and aligning video frames through a diffusion process tailored for multi-frame consistency.

The module begins by encoding consecutive video frames (frame_i and frame_{i+1}) into latent representations using a VAE, as shown in Figure 2. These latent encodings (z_t, z_{t+1}) are perturbed with Gaussian noise at each diffusion timestep to simulate noisy latent states (noise_1 and noise_2): $z_t \sim \mathcal{N}(\mu, \sigma^2)$. μ and σ^2 is the mean and variance of the Gaussian noise.

In this model, noise_1 and noise_2 are applied separately to z_t and z_{t+1} to capture frame-specific variations such as lighting, texture, and motion, ensuring that each frame's unique features are preserved. By injecting independent noise distributions, the model prevents over-smoothing, retaining high-quality spatial details while maintaining temporal consistency. It also simulates realistic

frame-to-frame degradations, which enhances the modelâĂŹs ability to reconstruct natural and consistent videos during denoising.

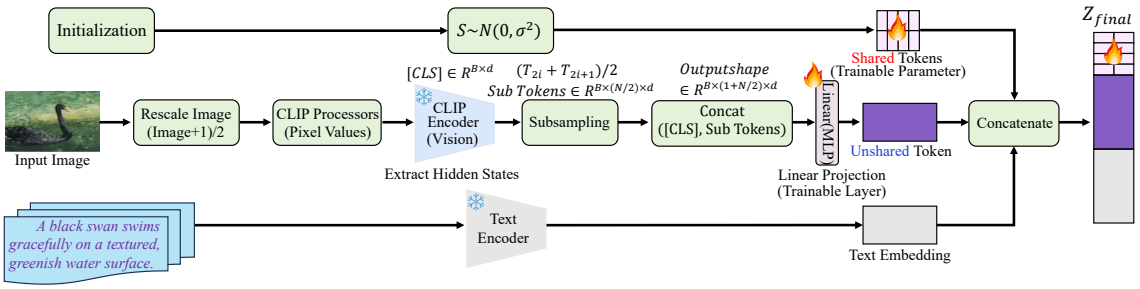

Figure 3: This method combines shared and unshared tokens, maintaining a consistent shared token embedding across timesteps while updating frame-specific unshared tokens. This balances global context and local details. Guided by text embeddings, the diffusion model refines noise into semantically consistent images. Frame-dependent unshared tokens introduce per-frame variations in cross-attention guidance and input noise.

The latent representations of consecutive frames, z_t and z_{t+1}, along with their noise-injected counterparts, are then combined into a unified latent representation $([z_t, z_{t+1}, z_t, z_{t+1}])$ to support cross-frame temporal modeling. This concatenated representation captures relationships between frames, enabling the model to jointly process both spatial and temporal features in the latent space. The noisy latent representations are further processed through a UNet-based architecture. Time embedding vectors, encoding the timestep information t, are introduced into the UNet to maintain structural consistency during the diffusion process. At each timestep, the latent representations are concatenated with control signals $(c_t, c_{t+1}, c_t, c_{t+1})$ generated by auxiliary ControlNet branches.

Latent representations for multiple frames are jointly processed through a combination of intermediate feature concatenation, time embedding integration, and latent concatenation. Temporal feature maps (F_t, F_{t+1}) are concatenated and injected into the UNet, enabling the model to capture inter-frame relationships effectively. Additionally, the timestep embeddings encoded within the UNet ensure that the noise schedule follows the temporal progression of the video. Furthermore, the concatenated latent representation $([z_t, z_{t+1}, z_t, z_{t+1}])$ enhances the modeling of spatial-temporal dependencies between frames.

3.2. Hierarchical-Aware Temporal-Spatial Consistency Module

Frame Similarity-based Temporal Consistency Blocks Diffusion-based video generation methods have advanced significantly in recent years, but maintaining temporal consistency across video frames remains a critical challenge. Existing models attempt to incorporate temporal dimensionality into the diffusion process using techniques such as pseudo-3D convolutions [74], sparse-causal attention [75], and self-attention feature injection [76]. However, these approaches often face drawbacks such as high computational costs, limited scalability, and suboptimal generalization to diverse video content. Unlike previous studies that primarily focus on large-scale retraining or post-hoc temporal smoothing, our approach introduces lightweight blocks with a novel temporal-aware loss function in the UNet's decoder layers. Details about these blocks are provided in Figure 4:

Further details on these blocks are shown in Figure 4:

$$x_{t+1} = x_t + \epsilon_t - \theta(x_t, t), \tag{1}$$

where x_t is the image at timestep t, ϵ_t is the predicted noise, and θ is the UNet model. This standard model produces static images and does not explicitly handle temporal relationships. We address this by incorporating lightweight, trainable adapters into the UNet, using hooks to extract intermediate feature maps $\mathbf{F}^t_{l,b}$ from each block (l, b) at timestep t:

$$\mathbf{F}^t_{l,b} = \mathbf{W}_0\mathbf{x} + \mathbf{B}_{l,b}\mathbf{A}_{l,b}\mathbf{x}, \tag{2}$$

where $\mathbf{x}$ is the input feature, and $\mathbf{B}_{l,b}$ and $\mathbf{A}_{l,b}$ are lightweight, learnable low-rank parameters.

To achieve smooth transitions between video frames, we use a similarity function to measure alignment between adjacent feature maps:

$$\text{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1}) = \frac{\mathbf{F}_t \cdot \mathbf{F}_{t+1}}{\|\mathbf{F}_t\| \|\mathbf{F}_{t+1}\|}. \tag{3}$$

We then define a temporal consistency loss:

$$\mathcal{L}_{\text{temporal}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \Big(\text{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1}) - \text{Sim}(\mathbf{F}_{t-1}, \mathbf{F}_t) \Big)^2, \tag{4}$$

where T is the total number of timesteps. By minimizing $\mathcal{L}_{\text{temporal}}$, the model aligns features between consecutive frames, reducing flickering and ensuring smoother transitions.

A standard diffusion loss is also required:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{x_0, \epsilon, t} \Big[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \Big], \tag{5}$$

where ϵ is the noise added to x_0 at timestep t, and ϵ_θ is the modelâĂŹs predicted noise. The overall objective function combines the temporal consistency and diffusion losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{temporal}} \mathcal{L}_{\text{temporal}} + \lambda_{\text{diffusion}} \mathcal{L}_{\text{diffusion}}, \tag{6}$$

where $\lambda_{\text{temporal}}$ and $\lambda_{\text{diffusion}}$ are set to 1 and 0.01, respectively. This balance promotes smooth, temporally consistent video editing while preserving denoising quality.

Channel-Dependent Spatially Consistent Denoising Blocks Another critical challenge in video editing is the inversion process, which is often necessary for generating edited outputs. Traditional frame-by-frame DDIM inversion [33, 77, 78] typically lack video-specific optimization, leading to inconsistencies across frames. To address this, we propose a bilateral filtering DDIM inversion technique that stabilizes latent representations and smooths spatial noisy latents without additional training, significantly improving frame-to-frame spatial consistency and ensuring seamless video generation.

In Denoising Diffusion Implicit Models (DDIM) inversion for video generation, preserving consistency and quality across consecutive frames is difficult because of the inherent noise variations in the diffusion process. The reverse diffusion process denoises an input x_t according to:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(t)\big), \tag{7}$$

where $\mu_\theta(x_t, t)$ is the predicted mean, and $\Sigma_\theta(t)$ is the variance schedule that manages uncertainty at each reverse step.

Existing DDIM-based video inversion techniques face significant challenges in achieving frame-level smoothness and quality. The stochastic nature of the diffusion process often introduces uneven textures, noise artifacts, and loss of details, leading to visually inconsistent frames. These issues arise from the probabilistic framework of the reverse diffusion process, where the denoising of a noisy input x_t is governed by:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sqrt{1-\alpha_{t-1}} z, \tag{8}$$

While $\epsilon_\theta(x_t, t)$ predicts the noise, and α_t, $\bar{\alpha}_t$ are scaling factors with z as sampled noise. We improve this framework with a bilateral filtering step applied to the noisy latents x_t. It reduces artifacts by preserving edges and smoothing noise based on spatial and intensity features:

$$O_x = \frac{\sum_{y \in \mathcal{N}(x)} G_{\text{spatial}}(x, y) G_{\text{intensity}}(I_x, I_y) I_y}{\sum_{y \in \mathcal{N}(x)} G_{\text{spatial}}(x, y) G_{\text{intensity}}(I_x, I_y)}, \tag{9}$$

where $\mathcal{N}(x)$ denotes the neighborhood of pixel x, with y as neighboring pixels contributing to smoothing based on spatial proximity and intensity similarity, defined by their respective intensities I_x and I_y. The spatial and intensity weights are calculated by:

$$G_{\text{spatial}}(x, y) = \exp\left(\frac{-(x-y)^2}{2\sigma^2_{\text{spatial}}} \right), \tag{10}$$

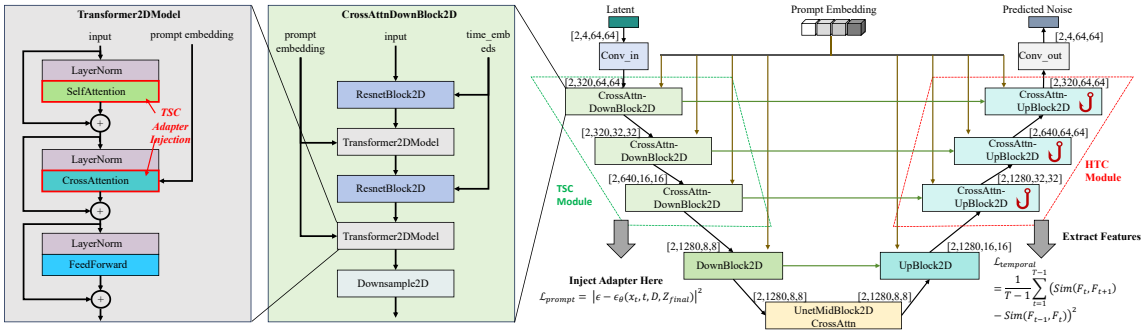

Figure 4: Overview of the UNet framework, showing where TSC adapters and HTC modules are introduced for temporal coherence and feature extraction. The CrossAttnDownBlock2D and UpBlock2D use prompt embeddings and time embeddings, with adapters added for better prompt conditioning. The losses L_{prompt} and L_{temporal} guide robust noise prediction.

$$G_{\text{intensity}}(I_x, I_y) = \exp\left(\frac{-(I_x - I_y)^2}{2\sigma^2_{\text{intensity}}}\right), \tag{11}$$

where σ_{spatial} determines sensitivity to spatial distances, and $\sigma_{\text{intensity}}$ controls the filter's response to intensity differences.

By incorporating bilateral filtering into the DDIM video inversion framework, the noisy latents x_t are smoothed at each timestep, producing refined latents x'_t. The updated inversion step is:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x'_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x'_t, t)\right) + \sqrt{1-\alpha_{t-1}}z, \tag{12}$$

where x'_t is the filtered latent obtained from x_t, ensuring smoother and more consistent intensity distributions. By replacing the original noisy latent with x'_t, the denoising aligns with the smoothed latent distribution. This step significantly reduces noise artifacts and improves overall frame quality throughout the video inversion process.

3.3. Token-Based Semantic Consistency Module

Existing video editing algorithms often encounter difficulties in aligning video frames with text semantics because of static embeddings and fragmented text integration [74–76], frequently producing semantic inconsistencies and visual artifacts such as flickering. To address these issues, we propose a Token-Based Semantic Consistency Module that merges shared tokens for global context alignment with dynamic unshared tokens for frame-specific details. The shared tokens ensure that the main semantics of the text prompt remain consistent across all frames, maintaining a coherent narrative throughout the video. Meanwhile, the dynamic unshared tokens adapt to frame-specific variations, capturing localized details such as texture, lighting, or motion, which are crucial for preserving temporal continuity. This combination allows our module to balance global coherence with per-frame adaptability, effectively reducing artifacts like flickering and enhancing the overall quality of video editing outputs. The shared token embeddings $T_{\text{share}} \in \mathbb{R}^{N_{\text{share}} \times 768}$ are initialized from a normal distribution $\mathcal{N}(0, 0.02)$, more details see Figure 3. For a given input image I, the CLIP model's vision encoder extracts visual hidden features $H_{\text{vision}} \in \mathbb{R}^{B \times N \times d}$, where B is the batch size, $N = 50$ is the sequence length, and $d = 768$ is the feature dimension. To construct a non-shared subset H_{sub}, adjacent feature vectors along the second dimension are averaged, resulting in $H_{\text{sub}} \in \mathbb{R}^{B \times N/2 \times d}$, defined as:

$$H_{\text{sub}}[:, i, :] = \frac{H_{\text{vision}}[:, 2i, :] + H_{\text{vision}}[:, 2i+1, :]}{2}, \quad \text{for } i \in \{0, 1, \ldots, N/2 - 1\}. \tag{13}$$

Next, a projection matrix $W \in \mathbb{R}^{d \times d}$ is applied to H_{sub} to preserve the feature dimension:

$$Z_{\text{unshare}} = H_{\text{sub}} \cdot W, \quad W = \text{nn.Linear}(d, d). \tag{14}$$

The final text embedding for temporal-aware fine-tuning is constructed as:

$$Z_{\text{final}} = [T_{\text{share}}; Z_{\text{frame}}; \mathcal{C}(Z)], \tag{15}$$

where T_{share} represents the shared token embedding, Z_{frame} is the frame-specific unshared token, and $\mathcal{C}(Z)$ concatenates conditional and unconditional embeddings along the first sequence dimension. The concatenation is denoted by $[\cdot]$.

During the denoising process, cross-attention provides text guidance by mapping the latent features $X_t \in \mathbb{R}^{M\times d}$ to updated features $\tilde{X}_t$ using the final text embedding $Z_{\text{final}} \in \mathbb{R}^{L\times d}$ as keys and values:

$$Q = W_Q^\top X_t, \quad K = W_K^\top Z_{\text{final}}, \quad V = W_V^\top Z_{\text{final}}, \tag{16}$$

$$\tilde{X}_t = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V. \tag{17}$$

where $W_Q \in \mathbb{R}^{M\times M}$, $W_K \in \mathbb{R}^{L\times d}$, and $W_V \in \mathbb{R}^{L\times d}$ are the learnable projection matrices for the query, key, and value transformations, respectively.

The updated cross-attention map $\tilde{X}_t$ is integrated into the noise prediction function ϵ_θ to guide the denoising process. The denoising step at timestep t can then be expressed as:

$$x_{t-1} = x_t - \alpha_t \epsilon_\theta(x_t, \tilde{X}_t), \tag{18}$$

where $x_t \in \mathbb{R}^{M\times d}$ (consistent with X_t and $\tilde{X}_t$) is the noisy latent at step t. The final text embedding Z_{final} integrates shared, frame-specific, and conditional/unconditional embeddings to compute $\tilde{X}_t$, aligning the denoising operation with frame-specific semantics to ensure global consistency and preserve local details.

During training, the parameters of the CLIP vision encoder (θ) remain frozen, while the adapter parameters, shared embeddings, and projection layers for unshared tokens (ϕ) are optimized iteratively to minimize the following loss function:

$$\text{Loss} = \begin{cases} |\epsilon - \epsilon_\theta(x_t, t, D, Z_{\text{final}})|^2, & \text{for } t \in [0, 0.5T], \\ |\epsilon - \epsilon_\theta(x_t, t, D, Z_{\text{final}})|^2 + \lambda \mathcal{L}_{\text{temproal}}, & \text{for } t \in [0.5T, T] \end{cases} \tag{19}$$

where ϵ represents the noise at timestep t, x_t is the input state, and $\mathcal{L}_{\text{temproal aware}}$ is the adjacent frames constraint. The adapter's parameters are updated as follows:

$$\Theta_{k+1} = \Theta_k - \eta \nabla_\Theta \text{Loss}(\Theta_k) \tag{20}$$

With $\Theta = \{\phi_{\text{adapter}}, \phi_{\text{unshared}}, T_{\text{share}}\}$ representing the adapter, unshared token, and shared token embedding parameters, and η as the learning rate, the adapter (ϕ_{adapter}) remains active only during the extended training interval from 0.5 to 1.0, where it influences ∇_ΘLoss, otherwise remaining inactive.

4. Experiments

4.1. Implementation Details

We trained a Stable Diffusion v1.5-based model with an 860M-parameter UNet and a 123M-parameter text encoder. We integrated a ControlNet specialized for depth data, coupled with an adapter that included a jointly trained PrefixToken module to enhance video-prompt alignment. Training was conducted in mixed precision (fp16) with a learning rate of 3e-5 and an input frame resolution of 512ÃŮ512. The 1.3G-parameter ControlNet was trained for 20 hours using four RTX4090 GPUs. The PrefixToken module comprises 2.3M parameters, and when jointly trained with the UNet adapter, it requires 18 hours using one RTX4090 GPU. The same hyperparameters were employed during ControlNet/PrefixToken training. We used the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The LoRA update matrix had a dimension of 4. The learning rate was 3e-5, with 500 warm-up steps in the scheduler, and the gradient accumulation steps were 8.

UNet adapter. UNet adapter only tunes attentions layer in UNet downsampling block.The adapter

parameter size is 860 M. Unet adapter and prompt adapter train together, needs 18 hours with only one RTX4090 GPU.

Prompt adapter. The prompt learner contains a trainable sharetoken and layers that map image information to tokens. The adapter parameter size is 123 M.

Controlnet. Controlnet initialized from Unet. The model parameter size is 1.3G. It trained on MSRVTT dataset for 20 hours with four RTX4090 GPUs.

4.2. Dataset

We used the MSR-VTT dataset [79], which has 10,000 video clips across 20 categories and 20 English captions each, resulting in about 29,000 unique words. The dataset is split into 6,513 training, 497 validation, and 2,990 test clips. We modified captions using the OpenAI ChatGPT API to produce additional variations and employed a specialized DataLoader to batch adjacent frames in early denoising steps. The MP4 videos were converted to WebDataset format for faster training. During inference, adapter weights were loaded into the StableDiffusionPipeline and merged at a 50% ratio.

Table 1: Comparison of algorithms with and without temporal awareness on the MSR-VTT test set. Adapter parameters were applied during the 0.9–1.0 denoising interval. Text-to-video generation uses only text input, so FID is not defined.

Algorithm	Adapter	LPIPS ↓	CLIP ↑	FID ↓
Text2Video-Zero [17]	✓	**0.319**	**31.38**	–
	✗	0.402	28.82	–
TokenFlow [80]	✓	**0.129**	**29.15**	**143.21**
	✗	0.135	28.90	166.72
Vid2Vid [81]	✓	**0.253**	**32.78**	154.30
	✗	0.303	31.94	**149.40**
VidToMe [82]	✓	**0.122**	**29.94**	**143.33**
	✗	0.123	29.93	146.27

4.3. Main Results

Superior Performance Across Algorithms and Settings Table 3 shows that our adapter is effective with different base diffusion models (e.g., SD-XL) and ControlNet conditioning modes. For instance, with Canny Edge conditioning, FID improves from 343.21 to 338.78, and LPIPS from 0.729 to 0.725. With Human Pose conditioning, CLIP scores increase from 29.49 to 33.56, while FID improves from 365.91 to 362.83 and LPIPS from 0.763 to 0.746, indicating enhanced semantic alignment. With Depth Map conditioning, FID decreases from 343.33 to 339.10, LPIPS from 0.721 to 0.718, and CLIP rises from 29.89 to 31.67. These results demonstrate the adapterâĂŹs ability to improve visual quality, semantic alignment, and frame-to-frame coherence under diverse conditions.

Table 1 highlights the compatibility of our adapter with various T2I-based T2V algorithms, showcasing its ability to consistently enhance performance across diverse methods [17, 34, 82, 83] with and without adapter injection demonstrates notable improvements in perceptual quality and text-image alignment: Text2Video-Zero achieves smoother transitions and better prompt adherence with LPIPS reduced from 0.402 to 0.319 and CLIP increased from 28.82 to 31.38; TokenFlow improves motion and texture consistency with a FID drop from 166.72 to 143.21, alongside LPIPS and CLIP gains; Vid2Vid enhances motion alignment with LPIPS decreasing from 0.303 to 0.253 and CLIP increasing from 31.94 to 32.78; and VidToMe improves perceptual quality and semantic alignment with LPIPS reduced from 0.126 to 0.114 and CLIP rising from 25.00 to 28.12, indicating enhanced perceptual quality and semantic alignment.

5. Conclusion

We introduced a prompt-learning adapter, GE-Adapter, to improve temporal consistency and visual quality in text-guided video editing using pre-trained text-to-image diffusion models. Our plug-and-

play adapter integrates temporal, spatial, and semantic consistency with Baliteral DDIM inversion, reducing flickering and refining text-to-video alignment at minimal training cost. Additionally, it incorporates three key componentsâĂŤFrame-based Temporal Consistency Blocks (FTC Blocks), Channel-dependent Spatial Consistency Blocks (SCD Blocks), and a Token-based Semantic Consistency Module (TSC Module)âĂŤto enhance perceptual quality and text-image relevance. This approach is also compatible with diverse video editing systems.

References

[1] Xiangfei Qiu, Xiuwen Li, Ruiyang Pang, Zhicheng Pan, Xingjian Wu, Liu Yang, Jilin Hu, Yang Shu, Xuesong Lu, Chengcheng Yang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, and Bin Yang. Easytime: Time series forecasting made easy. In *ICDE*, 2025.

[2] Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *SIGKDD*, pages 1185–1196, 2025.

[3] Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. In *Proc. VLDB Endow.*, pages 2363–2377, 2024.

[4] Yiyi Tao, Yixian Shen, Hang Zhang, Yanxin Shen, Lun Wang, Chuanqi Shi, and Shaoshuai Du. Robustness of large language models against adversarial attacks. In *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, pages 182–185. IEEE, 2024.

[5] Shaoshuai Du, Yiyi Tao, Yixian Shen, Hang Zhang, Yanxin Shen, Xinyu Qiu, and Chuanqi Shi. Zero-shot end-to-end relation extraction in chinese: A comparative study of gemini, llama and chatgpt. *arXiv preprint arXiv:2502.05694*, 2025.

[6] Yixian Shen, Hang Zhang, Yanxin Shen, Lun Wang, Chuanqi Shi, Shaoshuai Du, and Yiyi Tao. Altgen: Ai-driven alt text generation for enhancing epub accessibility. *arXiv preprint arXiv:2501.00113*, 2024.

[7] Yuqing Wang and Xiao Yang. Research on enhancing cloud computing network security using artificial intelligence algorithms. *arXiv preprint arXiv:2502.17801*, 2025.

[8] Yuqing Wang and Xiao Yang. Design and implementation of a distributed security threat detection system integrating federated learning and multimodal llm. *arXiv preprint arXiv:2502.17763*, 2025.

[9] Haopeng Zhao, Zhichao Ma, Lipeng Liu, Yang Wang, Zheyu Zhang, and Hao Liu. Optimized path planning for logistics robots using ant colony algorithm under multiple constraints. *arXiv preprint arXiv:2504.05339*, 2025.

[10] Letian Xu, Hao Liu, Haopeng Zhao, Tianyao Zheng, Tongzhou Jiang, and Lipeng Liu. Autonomous navigation of unmanned vehicle through deep reinforcement learning. In *Proceedings of the 5th International Conference on Artificial Intelligence and Computer Engineering*, pages 480–484, 2024.

[11] Xiangrui Xu, Qiao Zhang, Rui Ning, Chunsheng Xin, and Hongyi Wu. Comet: A communication-efficient and performant approximation for private transformer inference. *arXiv preprint arXiv:2405.17485*, 2024.

[12] Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561*, 2022.

[13] Jiachen Zhong and Yiting Wang. Enhancing thyroid disease prediction using machine learning: A comparative study of ensemble models and class balancing techniques, 2025.

[14] Revolutionizing drug discovery: Integrating spatial transcriptomics with advanced computer vision techniques, 2025. URL https://openreview.net/forum?id=deaeHR737W.

[15] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models, 2024. URL https://arxiv.org/abs/2402.17177.

[16] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report, 2023. URL https://arxiv.org/abs/2308.06571.

[17] 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, 2023. URL https://arxiv.org/abs/2303.13439.

[18] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation, 2024. URL https://arxiv.org/abs/2401.12945.

[19] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation, 2023. URL https://arxiv.org/abs/2212.11565.

[20] Yi Xin, Siqi Luo, Xuyang Liu, Haodi Zhou, Xinyu Cheng, Christina E Lee, Junlong Du, Haozhe Wang, MingCai Chen, Ting Liu, et al. V-petl bench: A unified visual parameter-efficient transfer learning benchmark. *Advances in Neural Information Processing Systems*, 37:80522–80535, 2024.

[21] Yi Xin, Junlong Du, Qiang Wang, Zhiwen Lin, and Ke Yan. Vmt-adapter: Parameter-efficient transfer learning for multi-task dense scene understanding, 2024.

[22] Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, and Yuntao Du. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint arXiv:2402.02242*, 2024.

[23] Mingyang Yi, Aoxue Li, Yi Xin, and Zhenguo Li. Towards understanding the working mechanism of text-to-image diffusion model. *arXiv preprint arXiv:2405.15330*, 2024.

[24] Siqi Luo, Yi Xin, Yuntao Du, Zhongwei Wan, Tao Tan, Guangtao Zhai, and Xiaohong Liu. Enhancing test time adaptation with few-shot guidance. *arXiv preprint arXiv:2409.01341*, 2024.

[25] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers, 2021.

[26] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers, 2022.

[27] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos, 2024.

[28] Cong Wang, Kuan Tian, Yonghang Guan, Jun Zhang, Zhiwei Jiang, Fei Shen, Xiao Han, Qing Gu, and Wei Yang. Ensembling diffusion models via adaptive feature aggregation, 2024.

[29] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation, 2024.

[30] Jiahao Huo, Zhengyao Wang, Rui Zhao, Lijun Sun, and Fei Shen. Synthesizing high-quality construction segmentation datasets through pre-trained diffusion model, 2024.

[31] Ruipeng Xu, Fei Shen, Xu Xie, and Zongyi Li. Training-free diffusion models for content-style synthesis, 2024.

[32] 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, 2023.

[33] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation, 2023. URL https://arxiv.org/abs/2212.11565.

[34] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing, 2023.

[35] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.08453.

[36] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2023. URL https://arxiv.org/abs/2211.01324.

[37] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions, 2023. URL https://arxiv.org/abs/2205.08534.

[38] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis, 2023. URL https://arxiv.org/abs/2212.05032.

[39] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing, 2023.

[40] Fei Shen, Xiangbo Shu, Xiaoyu Du, and Jinhui Tang. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval, 2023.

[41] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/076ccd93ad68be51f23707988e934906-Paper.pdf.

[42] Fei Shen, Yi Xie, Jianqing Zhu, Xiaobin Zhu, and Huanqiang Zeng. Git: Graph interactive transformer for vehicle re-identification, 2023.

[43] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, June 2020.

[44] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets, 2022. URL https://arxiv.org/abs/2202.00273.

[45] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinghui Tang. Imagdressing-v1: Customizable virtual dressing, 2024.

[46] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions, 2018. URL https://arxiv.org/abs/1804.08264.

[47] Petru Soviany, Claudiu Ardei, Radu Tudor Ionescu, and Marius Leordeanu. Image difficulty curriculum for generative adversarial networks (cugan), 2019. URL https://arxiv.org/abs/1910.08967.

[48] Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang. Advancing pose-guided image synthesis with progressive conditional diffusion models, 2023.

[49] Fei Shen, Hu Ye, Sibo Liu, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang. Boosting consistency in story visualization with rich-contextual conditional diffusion models, 2024.

[50] Bo Gao, Junchi Ren, Fei Shen, Mengwan Wei, and Zijun Huang. Exploring warping-guided features via adaptive latent diffusion model for virtual try-on, 2024.

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

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

[53] Fei Shen and Jinhui Tang. Imagpose: A unified conditional framework for pose-guided person generation, 2024.

[54] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022.

[55] Jinseok Kim and Tae-Kyun Kim. Arbitrary-scale image generation and upsampling using latent diffusion model and implicit neural decoder, 2024.

[56] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising, 2023.

[57] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing, 2023.

[58] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing, 2023.

[59] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing, 2023.

[60] Hanshu Yan, Jun Hao Liew, Long Mai, Shanchuan Lin, and Jiashi Feng. Magicprop: Diffusion-based video editing via motion-aware appearance propagation, 2023.

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

[62] Ziyue Zhang, Mingbao Lin, Shuicheng Yan, and Rongrong Ji. Easyinv: Toward fast and better ddim inversion, 2024. URL https://arxiv.org/abs/2408.05159.

[63] Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising, 2024. URL https://arxiv.org/abs/2403.14602.

[64] Wonjun Kang, Kevin Galim, and Hyung Il Koo. Eta inversion: Designing an optimal eta function for diffusion-based real image editing, 2024. URL https://arxiv.org/abs/2403.09468.

[65] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing, 2023. URL https://arxiv.org/abs/2304.08465.

[66] Jin Liu, Huaibo Huang, Chao Jin, and Ran He. Portrait diffusion: Training-free face stylization with chain-of-painting, 2023. URL https://arxiv.org/abs/2312.02212.

[67] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019. URL https://arxiv.org/abs/1902.00751.

[68] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

[69] Alec Radford. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. OpenAI technical report.

[70] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

[71] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2305.16322.

[72] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations, 2024. URL https://arxiv.org/abs/2403.06951.

[73] Archibald Fraikin, Adrien Bennetot, and StÃĽphanie AllassonniÃĺre. T-rep: Representation learning for time series using time-embeddings, 2024. URL https://arxiv.org/abs/2310.04486.

[74] 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, 2022. URL https://arxiv.org/abs/2209.14792.

[75] Wen Wang, Yan Jiang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models, 2023.

[76] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion, 2023.

[77] Qi Qian, Haiyang Xu, Ming Yan, and Juhua Hu. Siminversion: A simple framework for inversion-based text-to-image editing, 2024. URL https://arxiv.org/abs/2409.10476.

[78] Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models, 2024. URL https://arxiv.org/abs/2402.14780.

[79] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language, 2016.

[80] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing, 2023. URL https://arxiv.org/abs/2307.10373.

[81] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis, 2018. URL https://arxiv.org/abs/1808.06601.

[82] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing, 2023. URL https://arxiv.org/abs/2312.10656.

[83] Ernie Chu, Tzuhsuan Huang, Shuo-Yen Lin, and Jun-Cheng Chen. Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance, 2024.

A. Ablation Study

Ablation of Token-Based Semantic Consistency Module. Our ablation study confirms that shared tokens are essential for maintaining global semantic alignment. When the adapter model (including shared tokens) is used, it outperforms the Base model by keeping frame-level semantic focus on key elements such as a âĂIJswanâĂİ or âĂIJwater,âĂİ thus reducing attention drift. This also enhances details, for instance in neck movements or water ripples, and provides smooth transitions that tie motion and ripple effects together for stronger semantic consistency. Removing shared tokens severely degrades semantic alignment, as shown in Figure 15 (in supplementary material), where TokenFlow's CLIP score drops from 29.4 to 20.1, FID increases from 135.7 to 387.8, and LPIPS rises from 0.14 to 0.17, disrupting global coherence and local detail quality, as further visualized in Figure 12 (in supplementary material), with the lion's shape and textures becoming unrecognizable.

Table 2: Ablation Study: Effects of Inv on Different Algorithms.

Algorithm	Bilateral Inv	FID ↓	CLIP ↑	LPIPS ↓
Tokenflow	✗	135.66	29.17	0.142
	✓	**122.52**	**29.35**	**0.136**
Vid2Vid	✗	387.76	20.08	0.210
	✓	**320.80**	**25.84**	**0.166**
VidToMe	✗	357.25	25.00	0.126
	✓	**143.72**	**28.12**	**0.114**
Text2Video	✗	–	29.10	0.347
	✓	–	**30.07**	**0.244**

Boosting unshared tokens helps preserve attributes of the original video but can limit the modelâĂŹs flexibility for novel textual prompts. For example, TokenFlow struggles with âĂIJA lion is walking on the grass,âĂİ and Text2Video with âĂIJA man is running,âĂİ when unshared tokens are too heavily weighted.

Figures 15 and 12 12 (all in supplementary material) underline that too many unshared tokens degrade both global semantics and frame transitions. TokenFlowâĂŹs CLIP score, for instance, declines to 27.8, while FID and LPIPS deteriorate. A suitable mix of shared and unshared tokens is thus necessary for both fine detail and broad semantic integrity. We also vary the number of shared tokens used in the Prompt Adapter; visual results in Figure 13 (in supplementary material) and metric outcomes indicate that 18 shared tokens offer an effective balance. Excessive shared tokens risk losing text-specific information.

The activation time period of the Unet adapter is crucial, as shown in Figure 11 (in supplementary material), which evaluates the effects of temporal-aware loss training ranges and adapter activation ranges during inference. While a 0.5-1.0 training range combined with a 0.5-1.0 inference range achieves the best metrics across CLIP, FID, and LPIPS, Figure 14 (in supplementary material) reveals trade-offs: the broader 0.5-1.0 range introduces excessive constraints, resulting in over-smoothed details and visual blurriness. To resolve this, a narrower 0.9-1.0 inference range was adopted while keeping the broader 0.5-1.0 training range, striking a balance between temporal consistency and sharp, clear visual outputs.

Ablation of Hierarchical-Aware Temporal-Spatial Consistency Module We examine the effect of bilateral filtering within the Channel-Dependent Spatially Consistent Denoising Blocks. Figure 7 (in supplementary material) demonstrates that bilateral filtering reduces blur and jitter in generated frames, improving realism for objects such as penguins and rabbits. In Table 2, TokenFlowâĂŹs FID improves from 135.66 to 122.52, CLIP rises from 29.17 to 29.35, and LPIPS drops from 0.142 to 0.136, showing tangible gains in perceptual quality and coherence.

We also explored the advantages of our inversion method at smaller step sizes, as shown in Figure 5(in supplementary material), particularly in the 3-5 step configuration. At these smaller step sizes, our

Table 3: An ablation study on pretrained ControlNet conditions (controlnet-canny, openpose, depth-sdxl-1.0) with adapter results (yellow) on MSR-VTT human-edited cases for Human Pose demonstrates compatibility and performance gains, with enabled components in green.

Canny Edge	Human Pose	Depth Map	FID ↓	CLIP ↑	LPIPS ↓
✓	✗	✗	343.21 (338.78)	29.15 (31.44)	0.729 (0.725)
✗	✓	✗	365.91 (362.83)	29.49 (33.56)	0.763 (0.746)
✗	✗	✓	343.33 (339.10)	29.89 (31.67)	0.721 (0.718)

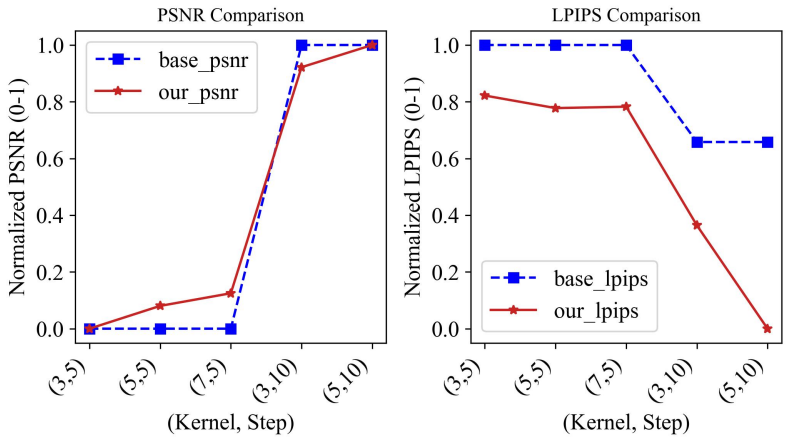

Figure 5: Comparison of step and kernel sizes $N(\cdot)$ within bilateral filtering inversion in the Stable Diffusion pipeline. The observed trends are consistent with those in vid2vid and three other algorithms (not shown due to space limitations).

inversion method significantly improves temporal coherence, resulting in smoother transitions between frames. This is especially crucial in video generation, where smaller step sizes reduce noise and enhance fine details by minimizing artifacts like flickering or blurry transitions that often occur at larger step sizes. As a result, the model demonstrates more accurate frame-to-frame consistency, with enhanced details and smoother motion, underscoring the effectiveness of our inversion method at smaller step sizes.

Furthermore, as is shown in Figure 5, we also observed that by appropriately increasing the kernel size (and its corresponding steps), image quality metrics (e.g., PSNR) can be enhanced while simultaneously improving coherence (e.g., LPIPS), ultimately producing clearer and more coherent visual outputs.

B. User Study

We selected 67 random participants with diverse genders, ages, and educational backgrounds. They were asked to evaluate video outputs based on three dimensions: the coherence between frames, the alignment between text and frames, and the quality of the video frames themselves. For each algorithm, we selected 10 video editing cases with an adapter and 10 without, forming 20 pairs of videos in total. Each pair included one video generated with the adapter and one without. The participants were unaware of which videos had the adapter applied and were informed only that the videos were generated using different algorithms. They were instructed to choose the video they considered the best in each pair. Afterward, we collected their preferences for videos with and without adapters across all four algorithms, along with their selections in the three evaluation dimensions of text-to-image alignment, image quality, and consistency. Finally, we averaged the data

from the 67 participants to calculate overall preference proportions for each algorithm and evaluation dimension. Figure 6 shows adapter-enhanced videos preferred across algorithms, with notable gains in consistency for VidToMe and image quality for TokenFlow, enhancing overall alignment and frame quality.

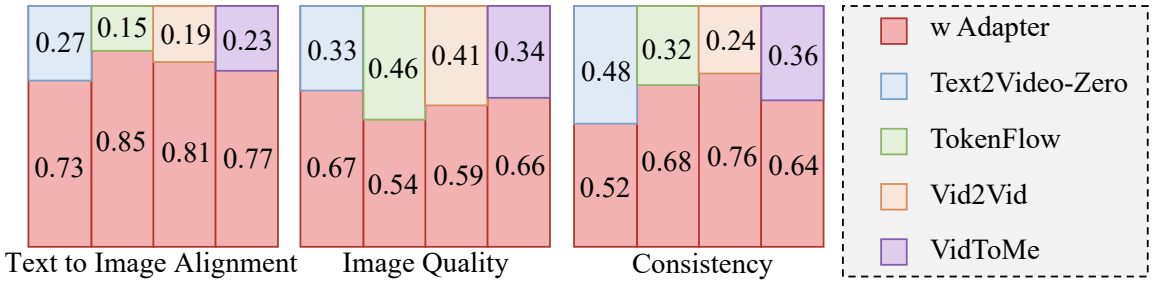

Figure 6: User study comparison of video generation algorithms with and without adding our adapter.

More experiment details and the qualitative results are in the supplementary material.

C. Theoretical Analysis

C.1. Optimizability of Temporal Consistency Loss

Theorem C.1 (Optimizability of Temporal Consistency Loss)**.** *Given A sequence of adjacent video frame feature maps* $\{\mathbf{F}_t\}_{t=1}^{T}$*, where* $\mathbf{F}_t \in \mathbb{R}^{H\times W\times C}$ *is the feature tensor of the t-th frame. The inter-frame similarity function:*

$$Sim(\mathbf{F}_t, \mathbf{F}_{t+1}) = \frac{\langle \mathbf{F}_t, \mathbf{F}_{t+1}\rangle}{\|\mathbf{F}_t\|_F \|\mathbf{F}_{t+1}\|_F}, \tag{21}$$

The temporal consistency loss $\mathcal{L}_{temporal}$*:*

$$\mathcal{L}_{temporal} = \frac{1}{T-1}\sum_{t=2}^{T-1} \left(Sim(\mathbf{F}_t, \mathbf{F}_{t+1}) - Sim(\mathbf{F}_{t-1}, \mathbf{F}_t)\right)^2. \tag{22}$$

If the norms of the feature maps are bounded (i.e., there exists $M > 0$ *such that* $\|\mathbf{F}_t\|_F \le M$ *for all* t*), then* $\mathcal{L}_{temporal}$ *is differentiable with respect to* $\{\mathbf{F}_t\}$ *and its gradient is Lipschitz continuous.*

To prove this theorem, we want to firstly prove the following lemma:

Lemma C.2 (Differentiability of the Cosine Similarity)**.** *For any* $\mathbf{F}_t, \mathbf{F}_{t+1}$*, the gradient of the cosine similarity:* $\nabla_{\mathbf{F}_t}\mathrm{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1})$ *Given the norm bound* $\|\mathbf{F}_t\|_F \le M$*, we have the bounded gradient:*

$$\|\nabla_{\mathbf{F}_t}\mathrm{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1})\|_F \le \frac{2}{M}. \tag{23}$$

Proof. For any two tensors: $\mathbf{F}_t, \mathbf{F}_{t+1} \in \mathbb{R}^{H\times W\times C}$with $\|\mathbf{F}_t\|_F \le M, \quad \|\mathbf{F}_{t+1}\|_F \le M, \quad$ for some $M > 0$,The cosine similarity as:
$\mathrm{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1}) := \frac{\langle \mathbf{F}_t, \mathbf{F}_{t+1}\rangle}{\|\mathbf{F}_t\|_F \|\mathbf{F}_{t+1}\|_F}$. So, its gradient with respect to $\mathbf{F}_t$ is:

$$\nabla_{\mathbf{F}_t}\mathrm{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1}) = \frac{\mathbf{F}_{t+1}}{\|\mathbf{F}_t\|_F\|\mathbf{F}_{t+1}\|_F} - \frac{\langle \mathbf{F}_t, \mathbf{F}_{t+1}\rangle}{\|\mathbf{F}_t\|_F^3\|\mathbf{F}_{t+1}\|_F}\mathbf{F}_t. \tag{24}$$

Using the submultiplicative property of the Frobenius norm, for the first term, we have

$$\left\|\frac{\mathbf{F}_{t+1}}{\|\mathbf{F}_t\|_F\|\mathbf{F}_{t+1}\|_F}\right\|_F = \frac{\|\mathbf{F}_{t+1}\|_F}{\|\mathbf{F}_t\|_F\|\mathbf{F}_{t+1}\|_F} = \frac{1}{\|\mathbf{F}_t\|_F}. \tag{25}$$

Since $\|\mathbf{F}_t\|_F \leq M$ the worstâĂŘcase (largest) value for the reciprocal is achieved when $\|\mathbf{F}_t\|_F$ is as small as possible; however, assuming that the features are nondegenerate (or alternatively invoking a lower bound implicitly provided by the normalization), we conclude that

$$\frac{1}{\|\mathbf{F}_t\|_F} \leq \frac{1}{M}. \tag{26}$$

Similarly, consider the second term:

$$\begin{aligned} \left\| \frac{\langle \mathbf{F}_t, \mathbf{F}_{t+1} \rangle}{\|\mathbf{F}_t\|_F^3 \|\mathbf{F}_{t+1}\|_F} \mathbf{F}_t \right\|_F &= \frac{|\langle \mathbf{F}_t, \mathbf{F}_{t+1} \rangle|}{\|\mathbf{F}_t\|_F^3 \|\mathbf{F}_{t+1}\|_F} \|\mathbf{F}_t\|_F \\ &= \frac{|\langle \mathbf{F}_t, \mathbf{F}_{t+1} \rangle|}{\|\mathbf{F}_t\|_F^2 \|\mathbf{F}_{t+1}\|_F}. \end{aligned} \tag{27}$$

By the CauchyâĂŞSchwarz inequality,

$$|\langle \mathbf{F}_t, \mathbf{F}_{t+1} \rangle| \leq \|\mathbf{F}_t\|_F \|\mathbf{F}_{t+1}\|_F, \tag{28}$$

and therefore

$$\frac{|\langle \mathbf{F}_t, \mathbf{F}_{t+1} \rangle|}{\|\mathbf{F}_t\|_F^2 \|\mathbf{F}_{t+1}\|_F} \leq \frac{\|\mathbf{F}_t\|_F \|\mathbf{F}_{t+1}\|_F}{\|\mathbf{F}_t\|_F^2 \|\mathbf{F}_{t+1}\|_F} = \frac{1}{\|\mathbf{F}_t\|_F}. \tag{29}$$

Again, using $\|\mathbf{F}_t\|_F \geq$ (a positive lower bound) and the worst case $\|\mathbf{F}_t\|_F \leq M$ we have

$$\frac{1}{\|\mathbf{F}_t\|_F} \leq \frac{1}{M}. \tag{30}$$

Combine the bounds, by the triangle inequality,

$$\|\nabla_{\mathbf{F}_t} \operatorname{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1})\|_F \leq \frac{1}{\|\mathbf{F}_t\|_F} + \frac{1}{\|\mathbf{F}_t\|_F} = \frac{2}{\|\mathbf{F}_t\|_F} \leq \frac{2}{M}. \tag{31}$$

This completes the proof of the differentiability and smoothness property for the cosine similarity function under the given norm boundedness assumption. Then we want to prove the next lemma:

Lemma C.3 (Lipschitz Continuity of the $\mathcal{L}_{\text{temporal}}$ Gradient)**.** *The gradient of $\mathcal{L}_{\text{temporal}}$ is: $\nabla_{\mathbf{F}_t} \mathcal{L}_{\text{temporal}}$ and $\Delta_t = \operatorname{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1}) - \operatorname{Sim}(\mathbf{F}_{t-1}, \mathbf{F}_t)$. Since $\operatorname{Sim}(\cdot, \cdot) \in [-1, 1]$, we have $|\Delta_t| \leq 2$. Combined with gradient boundedness, it follows that*

$$\left\| \nabla_{\mathbf{F}_t} \mathcal{L}_{\text{temporal}} \right\|_F \leq \frac{8}{M(T-1)}(T-2). \tag{32}$$

Thus, the gradient of $\mathcal{L}_{\text{temporal}}$ is Lipschitz continuous with constant $L \leq \frac{16}{M}$.

Proof. For clarity, we first state the expression for the gradient (with respect to $\mathbf{F}_t$) of $\mathcal{L}_{\text{temporal}}$:

$$\nabla_{\mathbf{F}_t} \mathcal{L}_{\text{temporal}} = \frac{2}{T-1} \sum_{t'=2}^{T-1} \Delta_{t'} \cdot \Big(\nabla_{\mathbf{F}_t} \operatorname{Sim}(\mathbf{F}_{t'}, \mathbf{F}_{t'+1}) - \nabla_{\mathbf{F}_t} \operatorname{Sim}(\mathbf{F}_{t'-1}, \mathbf{F}_{t'}) \Big), \tag{33}$$

Because the cosine similarity $\operatorname{Sim}(\cdot, \cdot)$ takes values in $[-1, 1]$ it follows immediately that $|\Delta_{t'}| \leq 2$. Moreover, from Lemma C.2 we have, for any pair $(\mathbf{F}, \mathbf{G})$:$\|\nabla_{\mathbf{F}} \operatorname{Sim}(\mathbf{F}, \mathbf{G})\|_F \leq \frac{2}{M}$. Thus, for any fixed index t and any summand in the expression, let

$$\psi_{t'} \coloneqq \nabla_{\mathbf{F}_t} \operatorname{Sim}(\mathbf{F}_{t'}, \mathbf{F}_{t'+1}) - \nabla_{\mathbf{F}_t} \operatorname{Sim}(\mathbf{F}_{t'-1}, \mathbf{F}_{t'}). \tag{34}$$

By the triangle inequality we have

$$\begin{aligned} \|\psi_{t'}\|_F &\leq \|\nabla_{\mathbf{F}_t} \operatorname{Sim}(\mathbf{F}_{t'}, \mathbf{F}_{t'+1})\|_F + \|\nabla_{\mathbf{F}_t} \operatorname{Sim}(\mathbf{F}_{t'-1}, \mathbf{F}_{t'})\|_F \\ &\leq \frac{2}{M} + \frac{2}{M} = \frac{4}{M}. \end{aligned} \tag{35}$$

Thus, for each t' we obtain

$$\|\Delta_{t'}\psi_{t'}\|_F \leq |\Delta_{t'}|\|\psi_{t'}\|_F \leq 2 \cdot \frac{4}{M} = \frac{16}{M}. \tag{36}$$

The overall gradient is given by averaging over the $T-2$ indices t' (from 2 to $T-1$). Hence, by the triangle inequality,

$$\begin{aligned} \|\nabla_{\mathbf{F}_t}\mathcal{L}_{\text{temporal}}\|_F &\leq \frac{2}{T-1}\sum_{t'=2}^{T-1}\|\Delta_{t'}\psi_{t'}\|_F \\ &\leq \frac{2}{T-1}(T-2)\cdot\frac{16}{M} \\ &= \frac{16(T-2)}{M(T-1)}. \end{aligned} \tag{37}$$

it follows immediately that

$$\|\nabla_{\mathbf{F}_t}\mathcal{L}_{\text{temporal}}\|_F \leq \frac{16}{M}\cdot\frac{T-2}{T-1} \leq \frac{16}{M}. \tag{38}$$

For any two admissible sets of feature tensors,

$$\begin{aligned} \|\nabla\mathcal{L}_{\text{temporal}}(\{\mathbf{F}_t\}) - \nabla\mathcal{L}_{\text{temporal}}(\{\mathbf{G}_t\})\| \leq L\cdot\sum_{t=1}^{T}\|\mathbf{F}_t - \mathbf{G}_t\| \\ \text{with } L \leq \frac{16}{M}. \end{aligned} \tag{39}$$

Thus, the gradient of $\mathcal{L}_{\text{temporal}}$ is Lipschitz continuous with Lipschitz constant

Theorem C.4 (Convergence of Gradient Descent)**.** *Let the parameters Θ be updated via gradient descent:*

$$\Theta_{k+1} = \Theta_k - \eta\nabla_\Theta \text{Loss}(\Theta_k), \quad \textit{(see Equation 20)} \tag{40}$$

where η is the learning rate. Suppose the gradient of $\mathcal{L}_{temporal}$ is L-Lipschitz continuous and $\eta < \frac{2}{L}$. Then $\mathcal{L}_{temporal}$ decreases monotonically and converges to a local minimum as $k \to \infty$.

To make a rigid proof of this theorem, we want to prove the following lemma first:

Lemma C.5 (Convexity of the Temporal Consistency Loss)**.** *Consider the temporal consistency loss $\mathcal{L}_{temporal}$ as a quadratic function of $\{\text{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1})\}_{t=1}^{T-1}$:*

$$\mathcal{L}_{temporal} = \frac{1}{T-1}\|\mathbf{D}\mathbf{s}\|_2^2, \tag{41}$$

where

$$\mathbf{s} = \left[\text{Sim}(\mathbf{F}_1, \mathbf{F}_2), \ldots, \text{Sim}(\mathbf{F}_{T-1}, \mathbf{F}_T)\right]^\top, \tag{42}$$

and $\mathbf{D}$ is the second-order difference matrix:

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(T-2)\times(T-1)}. \tag{43}$$

Since $\mathbf{D}^\top\mathbf{D}$ is positive semi-definite, $\mathcal{L}_{temporal}$ is convex with respect to the similarity terms $\text{Sim}(\mathbf{F}_t, \mathbf{F}_{t+1})$.

Proof. Express the loss as a quadratic form. Observe that

$$\begin{aligned} \mathcal{L}_{\text{temporal}} &= \frac{1}{T-1}\|\mathbf{D}\mathbf{s}\|_2^2 \\ &= \frac{1}{T-1}(\mathbf{D}\mathbf{s})^\top(\mathbf{D}\mathbf{s}) \\ &= \frac{1}{T-1}\mathbf{s}^\top\mathbf{D}^\top\mathbf{D}\mathbf{s}. \end{aligned} \tag{44}$$

For any vector $\mathbf{z} \in \mathbb{R}^{T-1}$, we have

$$\mathbf{z}^\top(\mathbf{D}^\top\mathbf{D})\mathbf{z} = (\mathbf{D}\mathbf{z})^\top(\mathbf{D}\mathbf{z}) = \|\mathbf{D}\mathbf{z}\|_2^2 \geq 0. \tag{45}$$

Thus, $\mathbf{D}^\top\mathbf{D}$ is indeed PSD, $\mathcal{L}_{\text{temporal}}$ is convex since $\mathbf{D}^\top\mathbf{D}$ is positive semidefinite.

Then we want to make a formal prove for C.4

Proof. By Lemma C.3 For all Θ, Θ' we have

$$\|\nabla\mathcal{L}_{\text{temporal}}(\Theta') - \nabla\mathcal{L}_{\text{temporal}}(\Theta)\|_2 \leq L\|\Theta' - \Theta\|_2. \tag{46}$$

By Lemma C.5, consequently, for any Θ, Θ',

$$\mathcal{L}_{\text{temporal}}(\Theta') \leq \mathcal{L}_{\text{temporal}}(\Theta) + \langle\nabla\mathcal{L}_{\text{temporal}}(\Theta), \Theta' - \Theta\rangle + \frac{L}{2}\|\Theta' - \Theta\|_2^2. \tag{47}$$

For the gradient descent update Equation 20, Set

$$\begin{aligned}\mathcal{L}_{\text{temporal}}(\Theta_{k+1}) \leq \mathcal{L}_{\text{temporal}}(\Theta_k) + \left\langle\nabla\mathcal{L}_{\text{temporal}}(\Theta_k), -\eta\nabla\mathcal{L}_{\text{temporal}}(\Theta_k)\right\rangle \\ + \frac{L}{2}\|\eta\nabla\mathcal{L}_{\text{temporal}}(\Theta_k)\|_2^2.\end{aligned} \tag{48}$$

The inner product term is

$$\left\langle\nabla\mathcal{L}_{\text{temporal}}(\Theta_k), -\eta\nabla\mathcal{L}_{\text{temporal}}(\Theta_k)\right\rangle = -\eta\|\nabla\mathcal{L}_{\text{temporal}}(\Theta_k)\|_2^2. \tag{49}$$

The squared norm is

$$\|\eta\nabla\mathcal{L}_{\text{temporal}}(\Theta_k)\|_2^2 = \eta^2\|\nabla\mathcal{L}_{\text{temporal}}(\Theta_k)\|_2^2. \tag{50}$$

Thus, we have:

$$\begin{aligned}\mathcal{L}_{\text{temporal}}(\Theta_{k+1}) &\leq \mathcal{L}_{\text{temporal}}(\Theta_k) - \eta\|\nabla\mathcal{L}_{\text{temporal}}(\Theta_k)\|_2^2 \\ &\quad + \frac{L\eta^2}{2}\|\nabla\mathcal{L}_{\text{temporal}}(\Theta_k)\|_2^2. \\ &\leq \mathcal{L}_{\text{temporal}}(\Theta_k) - \eta\left(1 - \frac{\eta L}{2}\right)\|\nabla\mathcal{L}_{\text{temporal}}(\Theta_k)\|_2^2.\end{aligned} \tag{51}$$

Note that since $0 < \eta < \frac{2}{L}$, the factor $\left(1 - \frac{\eta L}{2}\right)$ is positive. Hence, unless $\|\nabla\mathcal{L}_{\text{temporal}}(\Theta_k)\|_2^2 = 0$, the loss strictly decreases:

$$\mathcal{L}_{\text{temporal}}(\Theta_{k+1}) < \mathcal{L}_{\text{temporal}}(\Theta_k). \tag{52}$$

Since $\mathcal{L}_{\text{temporal}}$ is assumed to be bounded below, the sequence $\{\mathcal{L}_{\text{temporal}}(\Theta_k)\}$ is monotonically non-increasing and lower-bounded, and thus converges. Moreover, if the loss is convex, every stationary point is a global minimum. Hence, the iterates converge to a minimizer of $\mathcal{L}_{\text{temporal}}$.

By Theorem C.1, the temporal consistency loss $\mathcal{L}_{\text{temporal}}$ is differentiable, and its gradient is Lipschitz continuous once the feature maps $\{\mathbf{F}t\}$ are norm-bounded. This property guarantees that standard gradient-based methods can handle the optimization of $\mathcal{L}$temporalwithout diverging, because each gradient step remains well-controlled. Moreover, the Lipschitz condition implies that even in higher-dimensional latent spaces, the changes in the objectiveâĂŹs value do not fluctuate wildly with small alterations in the parameters.

In the accompanying corollary (not shown here but building on the same assumptions), one can establish that gradient descent converges to a local minimum under mild step-size requirements. This means the method has a sound mathematical basis for producing smooth transitions across video frames and for reducing flicker effects over time. From a practical standpoint, this reliability underpins the ability of the method to consistently refine temporal alignment, ensuring that each training iteration draws the system closer to a stable solution. Consequently, the theoretical analysis supports our claim that incorporating $\mathcal{L}_{\text{temporal}}$ leads to an approach that is both computable in practice and effective for generating temporally coherent video frames.

C.2. Stability of Bilateral Filtering DDIM Inversion

Theorem C.6 (Stability of Bilateral Filtering DDIM Inversion)**.** *Consider the DDIM inversion process (see Equation 12 in the main paper):*

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(x'_t - \frac{1-\alpha_t}{\sqrt{1-\tilde{\alpha}_t}}\epsilon_\theta(x'_t, t)\Big) + \sqrt{1-\alpha_{t-1}}z, \tag{53}$$

where x'_t is obtained by applying bilateral filtering (Equation 9) to the noisy latent x_t:

$$x'_t(y) = \frac{\sum_{x\in\mathcal{N}(y)} G_{\text{spatial}}(x,y)G_{\text{intensity}}(x_t(x), x_t(y))x_t(x)}{\sum_{x\in\mathcal{N}(y)} G_{\text{spatial}}(x,y)G_{\text{intensity}}(x_t(x), x_t(y))}. \tag{54}$$

Assume the following:

1. *The bilateral filter kernel parameters satisfy $\sigma_{\text{spatial}}, \sigma_{\text{intensity}} > 0$, and G_{spatial} and $G_{\text{intensity}}$ are Gaussian functions (Equations 10 and 11).*
2. *The noise predictor ϵ_θ is L_ϵ-Lipschitz continuous.*
3. *The ideal noise-free latent representation is $\bar{x}_t$, and the initial error satisfies $\mathbb{E}[\|x_T - \bar{x}_T\|_2] \le \delta$.*

Then there exists a constant $C = C(\alpha_t, \tilde{\alpha}_t, L_\epsilon) > 0$ such that the filtered latent representation satisfies

$$\mathbb{E}\big[\|x'_{t-1} - \bar{x}_{t-1}\|_2\big] \le C \cdot \mathbb{E}\big[\|x_t - \bar{x}_t\|_2\big] + \sqrt{1-\alpha_{t-1}}\mathbb{E}\big[\|z\|_2\big]. \tag{55}$$

We want to prove this theorem by proving three lemmas. Firstly, we want to prove:

Lemma C.7 (Error Contraction by Bilateral Filtering)**.** *Let $\mathcal{B}$ be the bilateral filtering operator mapping x_t to x'_t. By the weighted average property of bilateral filtering, we have*

$$\|x'_t - \bar{x}_t\|_2 = \Big\|\sum_x w(x,y)\big(x_t(x) - \bar{x}_t(x)\big)\Big\|_2 \tag{56}$$

where $w(x,y)$ are normalized weights. Since G_{spatial} and $G_{\text{intensity}}$ are exponentially decaying Gaussian functions, there exists $K > 0$ such that:

$$\|x'_t - \bar{x}_t\|_2 \le \|x_t - \bar{x}_t\|_2. \quad (1) \tag{57}$$

Hence, the bilateral filter is non-expansive

Proof. For each spatial location y, using the definition of the bilateral filtering operator, we have

$$x'_t(y) - \bar{x}_t(y) = \sum_{x\in\mathcal{N}(y)} w(x,y)\Big(x_t(x) - \bar{x}_t(x)\Big). \tag{58}$$

Taking the absolute value (or the norm in the scalar case) and applying the triangle inequality yields

$$\begin{aligned}\big|x'_t(y) - \bar{x}_t(y)\big| &= \left|\sum_{x\in\mathcal{N}(y)} w(x,y)\Big(x_t(x) - \bar{x}_t(x)\Big)\right| \\ &\le \sum_{x\in\mathcal{N}(y)} w(x,y)\big|x_t(x) - \bar{x}_t(x)\big|. \\ &\le \sup_{x\in\mathcal{N}(y)}\big|x_t(x) - \bar{x}_t(x)\big|.\end{aligned} \tag{59}$$

By the definition of the Euclidean norm, we have

$$\sup_{x\in\mathcal{N}(y)}\big|x_t(x) - \bar{x}_t(x)\big| \le \|x_t - \bar{x}_t\|_2. \tag{60}$$

Thus, for each y,

$$\big|x'_t(y) - \bar{x}_t(y)\big| \le \|x_t - \bar{x}_t\|_2. \tag{61}$$

Taking the L_2-norm over all spatial positions y on both sides, we obtain

$$\|x'_t - \bar{x}_t\|_2 \leq \|x_t - \bar{x}_t\|_2. \tag{62}$$

This establishes that the filtering operator $\mathcal{B}$ is non-expansive.

Lemma C.8 (Error Propagation in a Single DDIM Step)**.** *Consider the DDIM inversion process given by Equation 12 in the main paper. We decompose it into ideal and noisy paths:*

$$\bar{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(\bar{x}_t - \frac{1-\alpha_t}{\sqrt{1-\tilde{\alpha}_t}}\epsilon_\theta(\bar{x}_t, t)\Big), \tag{63}$$

Combining with the non-expansiveness result in Lemma C.7 gives:

$$\|x'_{t-1} - \bar{x}_{t-1}\|_2 \leq \underbrace{\Big(\frac{1}{\sqrt{\alpha_t}} + \frac{1-\alpha_t}{\sqrt{\alpha_t(1-\tilde{\alpha}_t)}}L_\epsilon\Big)}_{=:C}\|x_t - \bar{x}_t\|_2 + \sqrt{1-\alpha_{t-1}}\|z\|_2, \quad (2) \tag{64}$$

where C depends on α_t, $\tilde{\alpha}_t$, and L_ϵ.

Proof. Subtract the ideal inversion from the noisy one:

$$\begin{aligned} x'_{t-1} - \bar{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left[(x'_t - \bar{x}_t) - \frac{1-\alpha_t}{\sqrt{1-\tilde{\alpha}_t}}\Big(\epsilon_\theta(x'_t, t) - \epsilon_\theta(\bar{x}_t, t)\Big)\right] \\ + \sqrt{1-\alpha_{t-1}}z. \end{aligned} \tag{65}$$

Taking the L_2-norm and applying the triangle inequality gives:

$$\begin{aligned} \|x'_{t-1} - \bar{x}_{t-1}\|_2 \leq \frac{1}{\sqrt{\alpha_t}}\left(\|x'_t - \bar{x}_t\|_2 + \frac{1-\alpha_t}{\sqrt{1-\tilde{\alpha}_t}}\|\epsilon_\theta(x'_t, t) - \epsilon_\theta(\bar{x}_t, t)\|_2\right) \\ + \sqrt{1-\alpha_{t-1}}\|z\|_2. \end{aligned} \tag{66}$$

Since ϵ_θ is L_ϵ-Lipschitz, we have:

$$\|\epsilon_\theta(x'_t, t) - \epsilon_\theta(\bar{x}_t, t)\|_2 \leq L_\epsilon\|x'_t - \bar{x}_t\|_2. \tag{67}$$

Thus,

$$\|x'_{t-1} - \bar{x}_{t-1}\|_2 \leq \frac{1}{\sqrt{\alpha_t}}\left(1 + \frac{1-\alpha_t}{\sqrt{1-\tilde{\alpha}_t}}L_\epsilon\right)\|x'_t - \bar{x}_t\|_2 + \sqrt{1-\alpha_{t-1}}\|z\|_2. \tag{68}$$

By the result of the previous Lemma C.7 (non-expansiveness of the bilateral filtering operator),

$$\|x'_t - \bar{x}_t\|_2 \leq \|x_t - \bar{x}_t\|_2. \tag{69}$$

Substituting this into the previous inequality yields:

$$\|x'_{t-1} - \bar{x}_{t-1}\|_2 \leq \left(\frac{1}{\sqrt{\alpha_t}} + \frac{1-\alpha_t}{\sqrt{\alpha_t(1-\tilde{\alpha}_t)}}L_\epsilon\right)\|x_t - \bar{x}_t\|_2 + \sqrt{1-\alpha_{t-1}}\|z\|_2. \tag{70}$$

Defining

$$C \coloneqq \frac{1}{\sqrt{\alpha_t}} + \frac{1-\alpha_t}{\sqrt{\alpha_t(1-\tilde{\alpha}_t)}}L_\epsilon, \tag{71}$$

we obtain the desired bound:

$$\|x'_{t-1} - \bar{x}_{t-1}\|_2 \leq C\|x_t - \bar{x}_t\|_2 + \sqrt{1-\alpha_{t-1}}\|z\|_2. \tag{72}$$

Lemma C.9 (Expected Error Control in DDIM with Bilateral Filtering)**.** *From the single-step error bound in Lemma C.8), taking the expectation and using the independence assumption* $\mathbb{E}\big[\|z\|_2\big] = \sqrt{d}$ *(where* d *is the latent space dimension), we obtain*

$$\mathbb{E}\big[\|x'_{t-1} - \bar{x}_{t-1}\|_2\big] \leq C\mathbb{E}\big[\|x_t - \bar{x}_t\|_2\big] + \sqrt{1-\alpha_{t-1}}\sqrt{d}. \tag{73}$$

Recursively applying this from $t = T$ *down to* $t = 0$ *and using* $\mathbb{E}\big[\|x_T - \bar{x}_T\|_2\big] \leq \delta$ *yields*

$$\mathbb{E}\big[\|x'_0 - \bar{x}_0\|_2\big] \leq C^T\delta + \sqrt{d}\sum_{t=1}^{T} C^{t-1}\sqrt{1-\alpha_{t-1}}, \tag{74}$$

where $\alpha_t \in (0,1)$ *and* C *is the constant from the single-step analysis. Because* C *is bounded, the above series converges.*

Proof. Starting with the error propagation inequality,

$$\|x'_{t-1} - \bar{x}_{t-1}\|_2 \leq C\|x_t - \bar{x}_t\|_2 + \sqrt{1-\alpha_{t-1}}\|z\|_2, \tag{75}$$

we take expectations on both sides. Using the linearity of expectation and the independence of z, we obtain

$$\mathbb{E}\left[\|x'_{t-1} - \bar{x}_{t-1}\|_2\right] \leq C\mathbb{E}\left[\|x_t - \bar{x}_t\|_2\right] + \sqrt{1-\alpha_{t-1}}\sqrt{d}. \tag{76}$$

We now unroll inequality 76 recursively. Define $E_t := \mathbb{E}\left[\|x_t - \bar{x}_t\|_2\right]$. Then inequality 76 for time $t-1$ is

$$E_{t-1} \leq CE_t + \sqrt{1-\alpha_{t-1}}\sqrt{d}. \tag{77}$$

Applying this recursively from $t = T$ down to $t = 0$ proceeds as follows. For $t = T$:$E_T \leq \delta$.

For $t = T-1$:

$$\begin{aligned} E_{T-1} &\leq CE_T + \sqrt{1-\alpha_{T-1}}\sqrt{d} \\ &\leq C\delta + \sqrt{1-\alpha_{T-1}}\sqrt{d}. \end{aligned} \tag{78}$$

For $t = T-2$:

$$\begin{aligned} E_{T-2} &\leq CE_{T-1} + \sqrt{1-\alpha_{T-2}}\sqrt{d} \\ &\leq C\Big(C\delta + \sqrt{1-\alpha_{T-1}}\sqrt{d}\Big) + \sqrt{1-\alpha_{T-2}}\sqrt{d} \\ &= C^2\delta + C\sqrt{1-\alpha_{T-1}}\sqrt{d} + \sqrt{1-\alpha_{T-2}}\sqrt{d}. \end{aligned} \tag{79}$$

One obtains at $t = 0$:

$$E_0 = \mathbb{E}\left[\|x'_0 - \bar{x}_0\|_2\right] \leq C^T\delta + \sqrt{d}\sum_{t=1}^{T} C^{T-t}\sqrt{1-\alpha_{t-1}}. \tag{80}$$

Since $\alpha_t \in (0,1]$, for each t we have $\sqrt{1-\alpha_{t-1}} < 1$ and C is assumed bounded. Hence, the series

$$\sum_{t=1}^{T} C^{t-1}\sqrt{1-\alpha_{t-1}} \tag{81}$$

is a finite sum for fixed T and, when extended as $T \to \infty$ (if considering an infinite process), the bound remains meaningful provided that $C < 1$ or that other controlled conditions on the coefficients hold. In our case, for a fixed number of steps T, the series converges trivially as it is a finite sum.

This result tells us that the expected error at step $t-1$ is bounded by a constant C times the error at the previous step t plus an additional term that depends on the noise injection. If the constant C is less than or equal to oneâĂŤor even if it is slightly greater than one in the finite step caseâĂŤthe error term from the previous time-step does not get magnified significantly. Instead, it is either contracted or at most increased by a controlled constant factor. This behavior is often called an error contraction property.

The additional error that comes from the noise, given by $\sqrt{1-\alpha_{t-1}}\mathbb{E}[\|z\|_2]$, is also bounded (in many cases $\mathbb{E}[\|z\|_2] = \sqrt{d}$ where d is fixed). Therefore, at each step the noise adds a finite amount of error. When this recursive bound is applied over all steps (from the final time T to the initial time 0), the error at the final output is given by a geometric series-type bound (plus a sum of the noise contributions):

$$\mathbb{E}\big[\|x'_0 - \bar{x}_0\|_2\big] \leq C^T\delta + \sqrt{d}\sum_{t=1}^{T} C^{t-1}\sqrt{1-\alpha_{t-1}}, \tag{82}$$

where δ is the initial error at time T. Provided that C is bounded (and ideally $C < 1$ for true contraction), this series converges or remains finite for a finite number of steps.

Because neither the error propagation (scaled by C) nor the noise injection term causes the error to grow arbitrarily large during the reverse diffusion (DDIM inversion) process, the algorithm is stable. That is, the errors in the latent representation, when filtered and processed through each DDIM step, remain controlled.

In summary, the inequality shows that the DDIM inversion with bilateral filtering yields a bounded and controlled error propagation, thereby ensuring stability through the entire process.

C.3. Attention Alignment in Semantic Consistency Module) Statement

Theorem C.10 (Attention Alignment in Semantic Consistency Module)**.** *Let:*

- $X_t \in \mathbb{R}^{M\times d}$ *be the latent representation of frame t, where* M *is the number of spatial positions and* d *is the feature dimension.*
- $T_{share} \in \mathbb{R}^{N_s\times d}$ (*shared tokens*) *and* $Z_{unshare} \in \mathbb{R}^{N_u\times d}$ (*unshared tokens*) *form the joint embedding* $Z_{final} = \big[T_{share}; Z_{unshare}; \mathcal{C}(Z)\big] \in \mathbb{R}^{L\times d}$, *where* $L = N_s + N_u + \dim\big(\mathcal{C}(Z)\big)$.
- *There exist* $X^* \in \mathbb{R}^{M\times d}$ *and* $Z^* \in \mathbb{R}^{L\times d}$ *that achieve perfect semantic alignment:*

$$\begin{aligned} X^* &= \mathrm{softmax}\big(\tfrac{Q^*(K^*)^\top}{\sqrt{d}}\big)V^*, \\ &\text{where } Q^* = X^*W_Q,\ K^* = Z^*W_K,\ V^* = Z^*W_V. \end{aligned} \tag{83}$$

If the following conditions hold:

- *Full-rank projections:* W_Q, W_K, W_V *are invertible, and* $\sigma_{\min}(W_V) \geq \delta > 0$.
- *Token dimension sufficiency:* $N_s \geq d$ *and* $N_u \geq d$.
- *Lipschitz continuity: The Lipschitz constant of* softmax *satisfies* $L_{\mathrm{softmax}} \leq \sqrt{d}$.

Then the cross-attention output $\tilde{X}_t$ *in Equation 17 satisfies the alignment error bound:*

$$\begin{aligned} \|\tilde{X}_t - X^*\|_F &\leq \gamma\|Z_{final} - Z^*\|_F, \\ \gamma &= L_{\mathrm{softmax}}\tfrac{\|W_K\|_2\|W_V\|_2}{\delta}, \end{aligned} \tag{84}$$

where $\|\cdot\|_F$ *denotes the Frobenius norm.*

We plan to provide the proof in four lemmas.

Lemma C.11 (Decomposition of Cross-Attention)**.** *From Equations* (17) *and* (18)*, the attention output is given by*

$$\tilde{X}_t = \mathrm{softmax}\Big(\frac{X_tW_QW_K^\top Z_{\mathrm{final}}^\top}{\sqrt{d}}\Big)Z_{\mathrm{final}}W_V. \tag{85}$$

Define $\Delta Z = Z_{\text{final}} - Z^*$. The difference from the ideal output X^* can be decomposed into two terms:

$$\begin{aligned}\tilde{X}_t - X^* = &\underbrace{\Big(\text{softmax}\big(\tfrac{QK^\top}{\sqrt{d}}\big) - \text{softmax}\big(\tfrac{Q^*(K^*)^\top}{\sqrt{d}}\big)\Big)V^*}_{\textit{Term A}} \\ &+ \underbrace{\text{softmax}\big(\tfrac{QK^\top}{\sqrt{d}}\big)\Delta Z W_V}_{\textit{Term B}},\end{aligned} \tag{86}$$

where $Q = X_t W_Q$, $K = Z_{\text{final}} W_K$, $Q^* = X^* W_Q$, $K^* = Z^* W_K$, and $V^* = Z^* W_V$. Term A captures the discrepancy in attention weights, while Term B reflects the contribution of the adapter-induced change ΔZ.

Proof. We begin with the cross-attention output defined by 17

$$\tilde{X}_t = \text{softmax}\left(\frac{X_t W_Q W_K^\top Z_{\text{final}}^\top}{\sqrt{d}}\right) Z_{\text{final}} W_V. \tag{87}$$

Recall the following definitions:

$$Q = X_t W_Q, \quad K = Z_{\text{final}} W_K, \quad V = Z_{\text{final}} W_V, \tag{88}$$

and the ideal (perfectly aligned) quantities

$$\begin{aligned}&X^* = \text{softmax}\left(\frac{Q^*(K^*)^\top}{\sqrt{d}}\right) V^*, \\ &\text{where} \quad Q^* = X^* W_Q, \quad K^* = Z^* W_K, \quad V^* = Z^* W_V.\end{aligned} \tag{89}$$

Define the token embedding error as

$$\Delta Z = Z_{\text{final}} - Z^*. \tag{90}$$

Then note that the error in the value term is

$$\begin{aligned}V - V^* &= Z_{\text{final}} W_V - Z^* W_V \\ &= (Z_{\text{final}} - Z^*) W_V \\ &= \Delta Z W_V.\end{aligned} \tag{91}$$

Our goal is to show that

$$\begin{aligned}\tilde{X}_t - X^* = &\underbrace{\left(\text{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big) - \text{softmax}\Big(\frac{Q^*(K^*)^\top}{\sqrt{d}}\Big)\right) V^*}_{\text{Term A}} + \\ &\underbrace{\text{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)\Delta Z W_V}_{\text{Term B}}.\end{aligned} \tag{92}$$

To prove this, start by writing the expression for $\tilde{X}_t - X^*$:

$$\tilde{X}_t - X^* = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V - \text{softmax}\left(\frac{Q^*(K^*)^\top}{\sqrt{d}}\right) V^*. \tag{93}$$

We now add and subtract the same intermediate term $\text{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)V^*$ to decompose the expression:

$$\begin{aligned}\tilde{X}_t - X^* = &\left\{\text{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)V - \text{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)V^*\right\} \\ &+ \left\{\text{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)V^* - \text{softmax}\Big(\frac{Q^*(K^*)^\top}{\sqrt{d}}\Big)V^*\right\}.\end{aligned} \tag{94}$$

Notice that the first grouped term is

$$\mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)(V - V^*) = \mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)(\Delta Z W_V), \tag{95}$$

which is exactly Term B.

The second term becomes

$$\left[\mathrm{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big) - \mathrm{softmax}\Big(\frac{Q^*(K^*)^\top}{\sqrt{d}}\Big)\right] V^*, \tag{96}$$

which is Term A.

This completes the rigid mathematical proof of the decomposition.

Lemma C.12 (Bounding Term A)**.** *Using the Lipschitz property of the softmax function, we have*

$$\|\mathrm{softmax}(A) - \mathrm{softmax}(B)\|_F \le L_{\mathrm{softmax}} \|A - B\|_F. \tag{97}$$

Let $A = \frac{QK^\top}{\sqrt{d}}$ *and* $B = \frac{Q^*(K^*)^\top}{\sqrt{d}}$*. Then,*

$$\|\textit{Term A}\|_F \le L_{\mathrm{softmax}} \|W_Q\|_2 \|W_K\|_2 \|V^*\|_2 \|\Delta Z\|_F. \tag{98}$$

Proof. By the Lipschitz property and sub-multiplicativity of norms,

$$\begin{aligned} \|\text{Term A}\|_F &\le \|\,\mathrm{softmax}(A) - \mathrm{softmax}(B)\|_F \|V^*\|_2 \\ &\le L_{\text{softmax}} \|A - B\|_F \|V^*\|_2. \end{aligned} \tag{99}$$

Next, we bound $\|A - B\|_F$. By the definitions

$$A - B = \frac{1}{\sqrt{d}}\Big(QK^\top - Q^*(K^*)^\top\Big). \tag{100}$$

We can expand the difference as

$$QK^\top - Q^*(K^*)^\top = \underbrace{(Q - Q^*)K^\top}_{\text{Term 1}} + \underbrace{Q^*(K - K^*)^\top}_{\text{Term 2}}. \tag{101}$$

However, in our formulation the projection matrices W_Q and W_K are fixed (pretrained), i.e.,

$$\Delta W_Q = W_Q - W_Q = 0, \qquad \Delta W_K = W_K - W_K = 0. \tag{102}$$

Since

$$Q = X_t W_Q \quad \text{and} \quad Q^* = X_t W_Q, \tag{103}$$

it follows that $Q - Q^* = 0$ so that Term 1 is identically zero. Next, observe that

$$K = Z_{\text{final}} W_K \quad \text{and} \quad K^* = Z^* W_K. \tag{104}$$

Hence,

$$K - K^* = (Z_{\text{final}} - Z^*) W_K = \Delta Z W_K. \tag{105}$$

Therefore, Term 2 becomes

$$Q^*(K - K^*)^\top = Q^*(W_K^\top \Delta Z^\top) = X_t W_Q W_K^\top \Delta Z^\top. \tag{106}$$

Gathering the above, we deduce

$$\|A - B\|_F = \frac{1}{\sqrt{d}} \Big\| X_t W_Q W_K^\top \Delta Z^\top \Big\|_F. \tag{107}$$

Using the sub-multiplicative property of the Frobenius norm and the fact that for any matrix X, $\|X\|_F \le \sqrt{r}\|X\|_2$ when r is the rank (or simply using the induced norm properties), we can bound

$$\|X_t W_Q W_K^\top \Delta Z^\top\|_F \le \|X_t\|_F \|W_Q\|_2 \|W_K\|_2 \|\Delta Z^\top\|_2. \tag{108}$$

Note that $\|\Delta Z^\top\|_2 = \|\Delta Z\|_2 \leq \|\Delta Z\|_F$ (since the spectral norm is bounded by the Frobenius norm). Thus, we have

$$\|A - B\|_F \leq \frac{\|W_Q\|_2 \|W_K\|_2}{\sqrt{d}} \|X_t\|_F \|\Delta Z\|_F. \tag{109}$$

Plugging this back into the bound for Term A, we obtain

$$\|\text{Term A}\|_F \leq L_{\text{softmax}} \frac{\|W_Q\|_2 \|W_K\|_2}{\sqrt{d}} \|X_t\|_F \|\Delta Z\|_F \|V^*\|_2. \tag{110}$$

In many applications the feature matrix X_t may be normalized such that $\|X_t\|_F \leq \sqrt{d}$ (or this factor can be absorbed into the Lipschitz constant or constant of proportionality). Under such a normalization, we arrive at the final bound

$$\|\text{Term A}\|_F \leq L_{\text{softmax}} \|W_Q\|_2 \|W_K\|_2 \|V^*\|_2 \|\Delta Z\|_F. \tag{111}$$

This completes the rigorous derivation of the bound on Term A.

Lemma C.13 (Bound on Term B)**.** *Under the normalization property of the softmax function* ($\|\text{softmax}(\cdot)\|_F \leq 1$) *and the full-rank condition on W_V, the following holds:*

$$\|\textit{Term B}\|_F \leq \|W_V\|_2 \|\Delta Z\|_F. \tag{112}$$

Proof. Since the softmax operator normalizes its input such that each row is a probability distribution, we have

$$\left\| \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} \right) \right\|_F \leq 1. \tag{113}$$

Thus, by the submultiplicativity of the Frobenius norm,

$$\|\text{Term B}\|_F \leq \|\Delta Z W_V\|_F. \tag{114}$$

Next, applying the standard inequality for matrix norms,

$$\|\Delta Z W_V\|_F \leq \|W_V\|_2 \|\Delta Z\|_F. \tag{115}$$

Therefore, we obtain the desired bound:

$$\|\text{Term B}\|_F \leq \|W_V\|_2 \|\Delta Z\|_F. \tag{116}$$

Lemma C.14 (Combined Error Bound on Cross-Attention)**.** *By merging Term A and Term B from the cross-attention decomposition, we have:*

$$\|\tilde{X}_t - X^*\|_F \leq \underbrace{\left(L_{\text{softmax}} C \|W_Q\|_2 \|W_K\|_2 \|W_V\|_2 + \|W_V\|_2 \right)}_{\gamma} \|\Delta Z\|_F. \tag{117}$$

Upon simplification, γ can be expressed as

$$\gamma = \frac{L_{\text{softmax}} \|W_K\|_2 \|W_V\|_2}{\delta} \, (\textit{absorbing } \|W_Q\|_2 \textit{ into constants}). \tag{118}$$

Proof. We know from the previous bounds that

By Lemma C.12, Term A satisfies

$$\|\text{Term A}\|_F \leq L_{\text{softmax}} \|W_Q\|_2 \|W_K\|_2 \|V^*\|_2 \|\Delta Z\|_F, \tag{119}$$

and by Lemma C.13, Term Bsatisfies

$$\|\text{Term B}\|_F \leq \|W_V\|_2 \|\Delta Z\|_F. \tag{120}$$

The triangle inequality implies

$$\|\tilde{X}_t - X^*\|_F \leq (L_{\text{softmax}}\|W_Q\|_2\|W_K\|_2\|V^*\|_2 + \|W_V\|_2)\,\|\Delta Z\|_F. \tag{121}$$

Recall that by definition $V^* = Z^*W_V$. Let us assume that the ideal token matrix is bounded, namely $\|Z^*\|_2 \leq C$. Then, by submultiplicativity of the spectral norm,

$$\|V^*\|_2 \leq \|Z^*\|_2\|W_V\|_2 \leq C\|W_V\|_2. \tag{122}$$

That is, defining

$$\gamma := L_{\text{softmax}}C\|W_Q\|_2\|W_K\|_2\|W_V\|_2 + \|W_V\|_2, \tag{123}$$

we have

$$\|\tilde{X}_t - X^*\|_F \leq \gamma\|\Delta Z\|_F. \tag{124}$$

Using the full-rank assumption that $\sigma_{\min}(W_V) \geq \delta > 0$,the smallest singular value of W_V is bounded away from zero. We can absorb $\|W_Q\|_2$ or other fixed constants into the constant. In fact, if we reparameterize or normalize the matrices appropriately, we may simplify the bound to:

$$\gamma = \frac{L_{\text{softmax}}\|W_K\|_2\|W_V\|_2}{\delta}, \tag{125}$$

by absorbing the constant C and $\|W_Q\|_2$ into δ (or equivalently assuming that the constant factors have been normalized).

Thus, the final combined error bound is

$$\begin{aligned}\|\tilde{X}_t - X^*\|_F &\leq (L_{\text{softmax}}\|W_Q\|_2\|W_K\|_2\|V^*\|_2 + \|W_V\|_2)\,\|\Delta Z\|_F \\ &\leq \frac{L_{\text{softmax}}\|W_K\|_2\|W_V\|_2}{\delta}\|\Delta Z\|_F.\end{aligned} \tag{126}$$

Corollary C.15 (Token Sufficiency)**.** *If $N_s \geq d$ and $N_u \geq d$, then the column space of Z_{final} spans $\mathbb{R}^d$, allowing $\|\Delta Z\|_F$ to be minimized through optimization. Consequently, the error bound $\gamma\|\Delta Z\|_F$ converges to zero, leading to exact semantic alignment.*

Theorem C.10 shows that the alignment error of the cross-attention output can be made arbitrarily small if the difference $\|\Delta Z\|_F$ between the learned and ideal token embeddings is reduced. The bound involves a constant γ that depends on the Lipschitz continuity of softmax and the singular values of the projection matrices. Corollary C.15 states that if the number of shared and unshared tokens meets or exceeds the feature dimension ($N_s \geq d$ and $N_u \geq d$), then the column space of the token embeddings spans all possible feature directions. In other words, the learned tokens can represent any point in $\mathbb{R}^d$, ensuring that $\|\Delta Z\|_F$ can be minimized through standard optimization techniques. As a result, the cross-attention module can achieve exact semantic alignment.

This theoretical result underpins the stability and effectiveness of the proposed method. By ensuring that the token embeddings are sufficiently rich, the attention alignment error can be driven to zero, which guarantees that semantic information is accurately preserved and transferred.

D. Qualitative Results

Figure 9 (in supplementary material) highlights the **crucial** role of the adapter in enhancing video quality across a range of scenarios and algorithms. In the âĂIJlion roaringâĂİ example, videos without the adapter show inconsistent facial features and unsteady motion, while the adapter achieves fluid transitions and a coherent depiction of the roaring action. In the âĂIJchild biking through waterâĂİ scenario, videos without the adapter suffer from artifacts and temporal inconsistencies, including distorted water reflections and unnatural motion. The adapter addresses these issues effectively, creating smooth biking dynamics and realistic water effects. Similarly, in the âĂIJwalking dogâĂİ example, frame flickering and uneven body movements occur without the adapter, but are eliminated when it is used, producing smoother, more natural strides. Finally, in the âĂIJman

surfingâĂİ illustration, the adapter strengthens semantic alignment by maintaining the surferâĂŹs posture and interaction with the waves, resulting in visually cohesive and dynamic transitions. These case studies demonstrate how the adapter can improve temporal stability and semantic alignment, ensuring high-quality video outputs.

In Figures 10 and 8 (all in supplementary material), we compare cross-attention maps generated by the base Stable Diffusion model and our adapter-enhanced model during the image-generation process. Minor differences in these attention maps underscore the adapterâĂŹs influence on the modelâĂŹs performance.

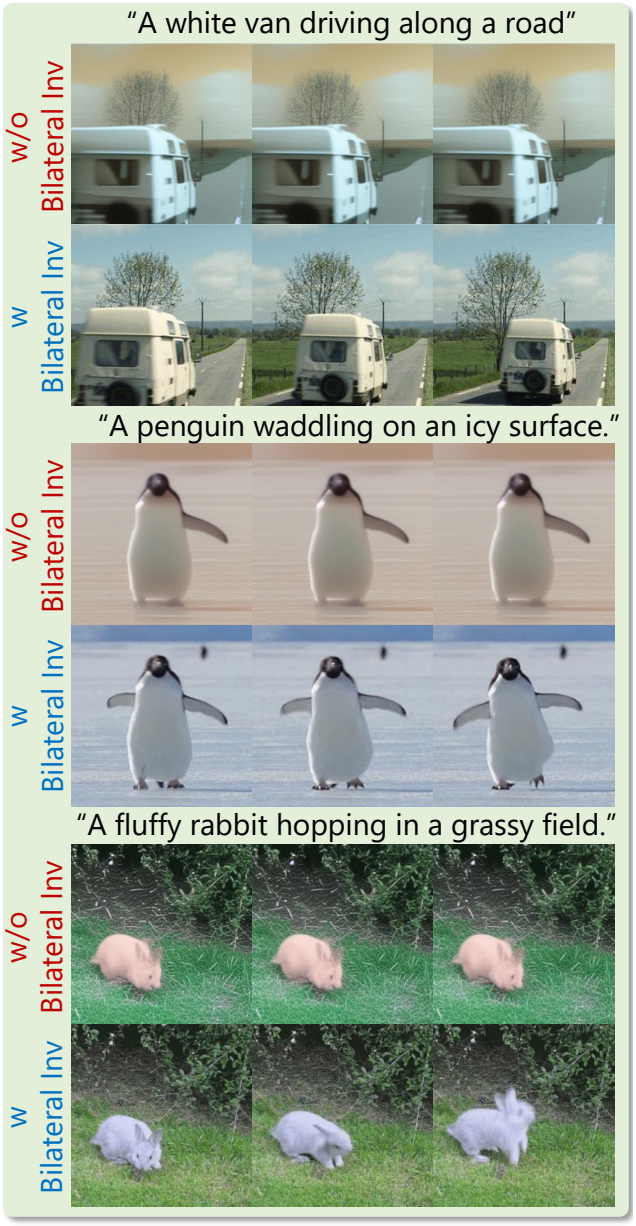

Figure 7: Comparison of video generation results with bilateral filtering using different kernel sizes in the Stable Diffusion pipline.

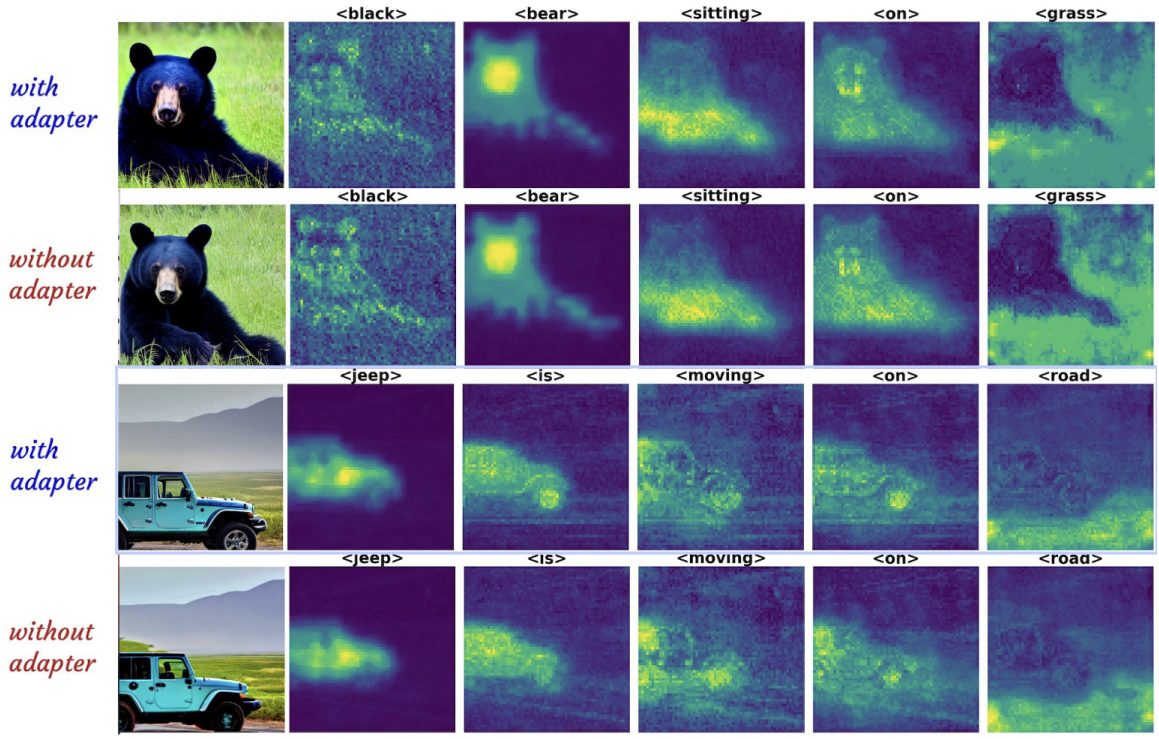

Figure 8: Visualization of attention maps comparing video frames generated with and without the adapter in the Stable Diffusion 1.5 pipeline.

Figure 9: A comparison of video generation quality across different algorithms, with and without the use of an adapter.

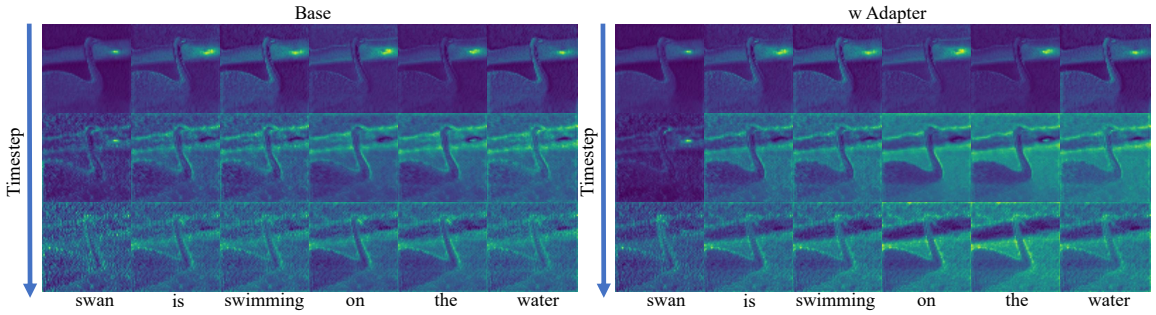

Figure 10: Visualization of attention maps from Unet's third upsampling block, comparing base and adapter models across 1000 timesteps. The corresponding timesteps from top to bottom are 1, 541, and 981.

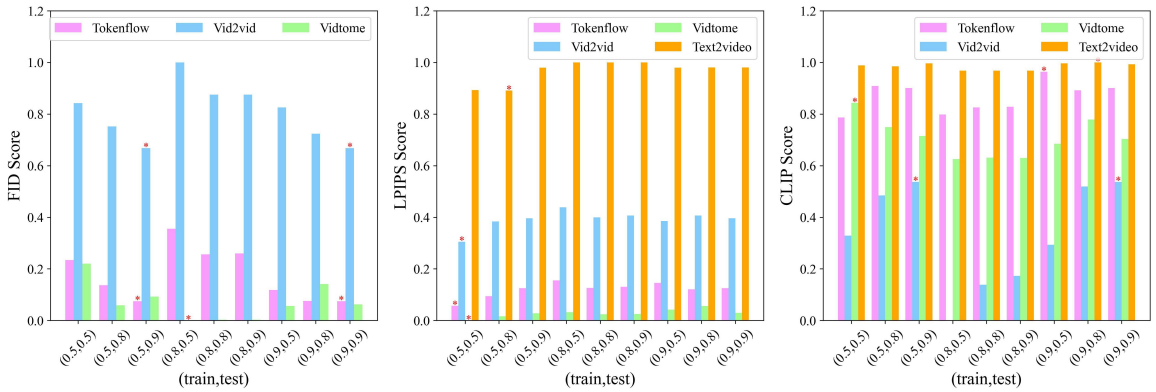

Figure 11: Bar chart illustrating the performance of four algorithms across three metrics for various combinations of (training timesteps, inference timesteps). The values 0.5, 0.8, and 0.9 represent the ranges 0.5tâĂŞ1.0t, 0.8tâĂŞ1.0t, and 0.9tâĂŞ1.0t, respectively, where t denotes 1000 timesteps. The number on the left of the tuple indicates the training timesteps, while the number on the right represents the inference timesteps. Note that Text2Video, as a text-to-video generation algorithm, lacks pre-edited videos and therefore does not have an FID metric. The best (training timesteps, inference timesteps) combination for each algorithm is marked with an asterisk (*) at the top of the corresponding bar.

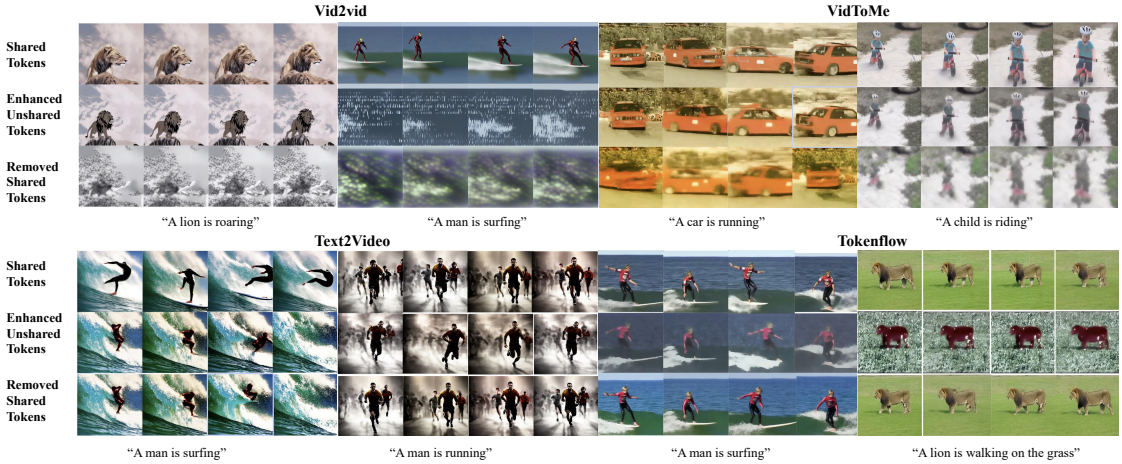

Figure 12: Visualization of video generation models (Vid2Vid, VidToMe, Text2Video, and TokenFlow) using different token strategies: Shared Tokens, Enhanced Unshared Tokens, and Removed Shared Tokens, evaluated across multiple prompts.

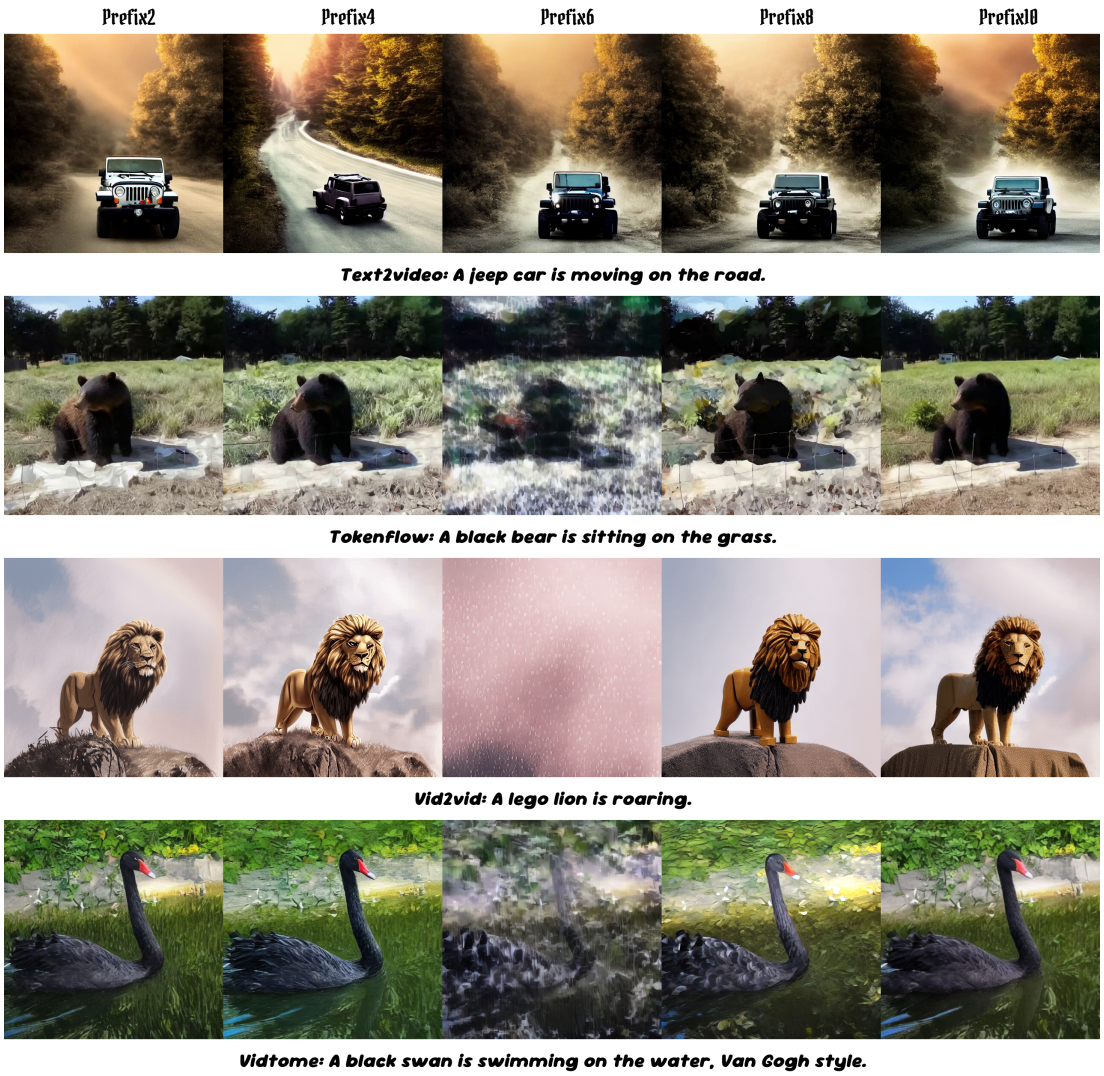

Figure 13: Visualization results obtained by Prompt learner taking different number of shared token (2, 4, 6, 8, 10) training and inference.

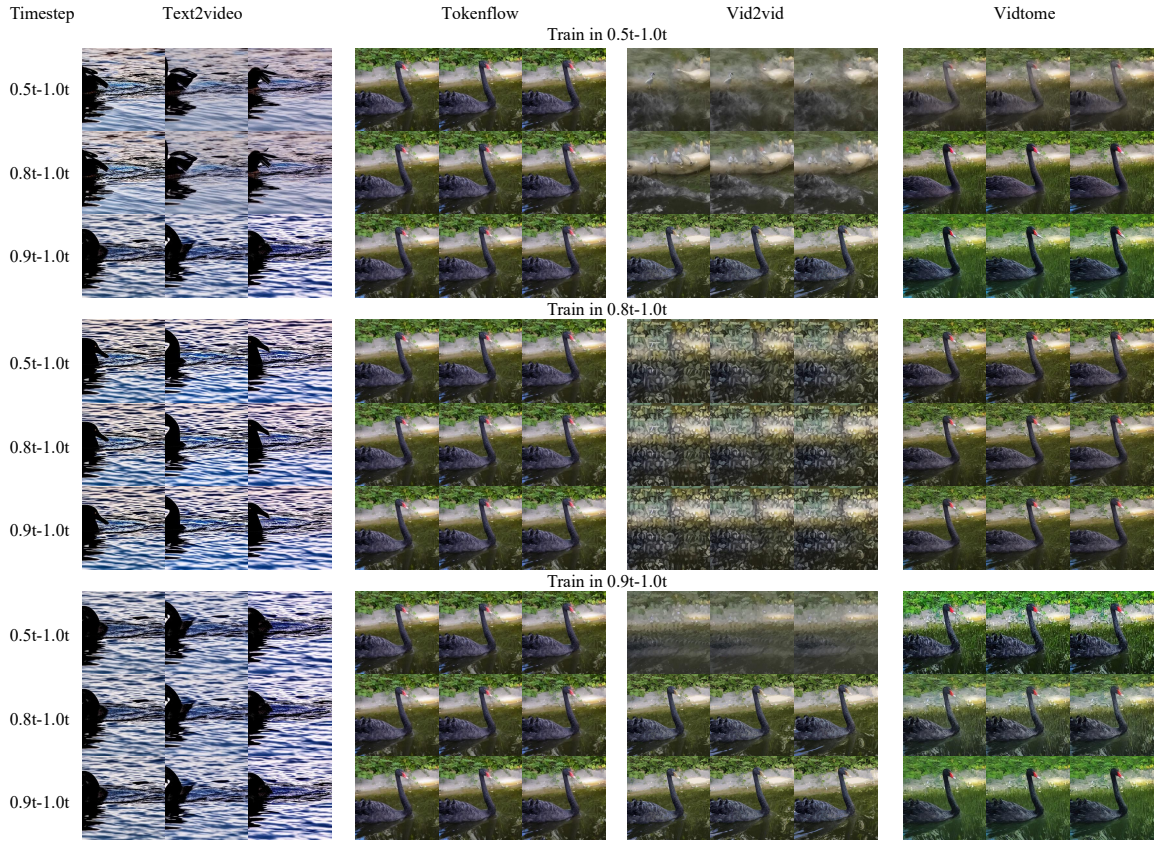

Figure 14: Visual comparison of a 0.5-1.0t training range for temporal-aware loss with a 0.9-1.0t adapter activation in inference range versus setting both to 0.5t (t = 1000).

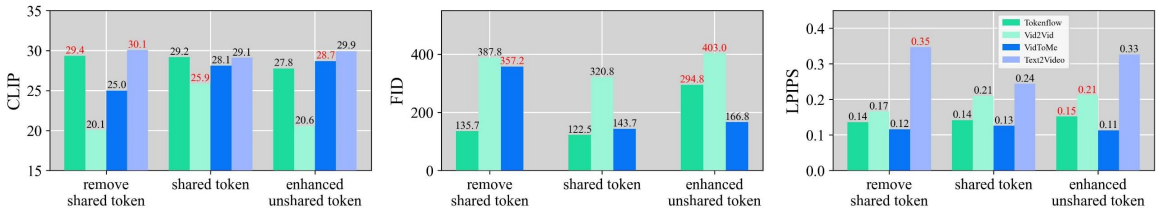

Figure 15: Comparison of token configurationsâĂŤremoving shared tokens, using shared tokens, and using enhanced unshared tokensâĂŤbased on LPIPS, CLIP, and FID metrics.