MagicPrompt: Ultra-Lightweight Prompt Tuning for Video Generation

Yinhan Zhang[1*], Dingwei Tan[2*], Xianghao Kong[1], Yue Ma[1‡], Yeying Jin[3✉], and Anyi Rao[1✉]

[1] The Hong Kong University of Science and Technology, Hong Kong
[2] The Hong Kong University of Science and Technology(Guangzhou), Guangzhou
[3] Tencent, Singapore

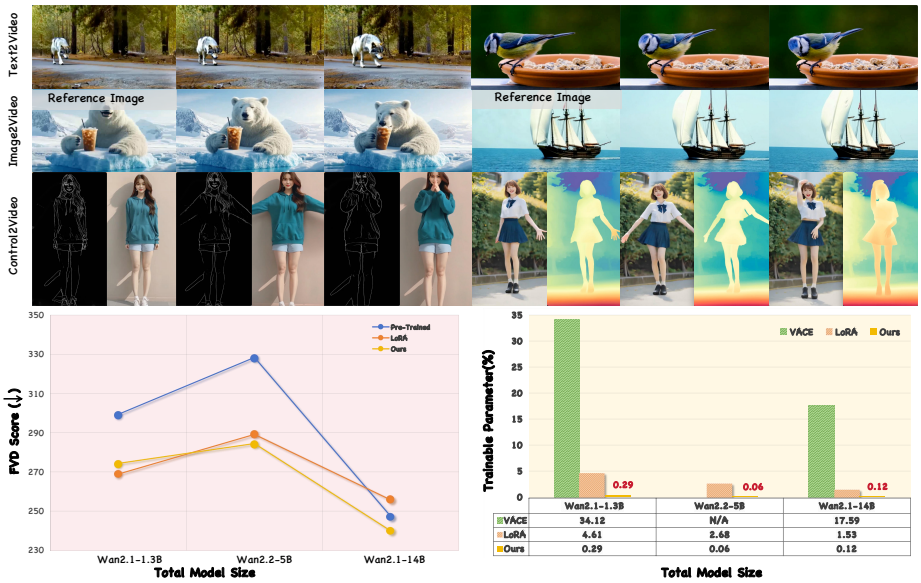

Fig. 1: Overview of MagicPrompt. We introduce MagicPrompt, an efficient attention-embedded prompt tuning framework for video diffusion models. Across diverse model sizes, MagicPrompt achieves competitive generation quality compared to Full Fine-tuning and LoRA, while requiring the fewest trainable parameters. Our method demonstrates seamless adaptation to Text-to-Video (T2V), Image-to-Video (I2V), and Control-to-Video (Control2V) tasks.

Abstract. Large-scale video diffusion models (VDMs) deliver strong generation performance, but full fine-tuning for downstream tasks incurs prohibitive computational costs. Existing parameter-efficient fine-tuning (PEFT) methods have two critical flaws on billion-scale models: they still require substantial trainable parameters, and reward-based

∗ Equal contribution; ‡ Project Leader; ✉ Corresponding author.

training suffers from noise-induced optimization instability in condition-guided tasks. We propose MagicPrompt, a lightweight framework that achieves extreme parameter efficiency and stable reward optimization. It first adopts Attention-Embedded Prompt Tuning, which steers generation via lightweight soft prompts with orders of magnitude fewer parameters while preserving pre-trained knowledge. It further introduces Dual-Space Reward Feedback Optimization, which uses self-supervised latent objectives to improve condition-guided reward training. Experiments show MagicPrompt reaches competitive performance with less than 1% trainable parameters and notably reduces training costs.

Keywords: Parameter Efficient Fine Tuning · Diffusion Model

1 Introduction

The advent of large-scale video diffusion models (VDMs) [23, 29, 30, 32–34, 71] has marked a paradigm shift in generative artificial intelligence, demonstrating remarkable capabilities in synthesizing high-quality, temporally coherent videos from textual prompts. However, the sheer scale of these models presents a significant barrier to practical adoption, as adapting a pre-trained VDM to downstream tasks typically requires full fine-tuning that updates billions of parameters. This approach is not only computationally prohibitive due to immense memory and storage demands but also risks destabilizing the model's original generative distribution. To mitigate these costs, **Parameter-Efficient Fine-Tuning (PEFT)** has emerged as a critical direction, aiming to customize massive models by updating only a minimal subset of parameters while freezing the pre-trained backbone. Despite this promise, adapting video diffusion models remains an open challenge characterized by interconnected bottlenecks: existing methods often suffer from residual parameter inefficiency and computational overhead, while still demanding large-scale datasets that are scarce for video tasks. More critically, under limited-data settings, these approaches are prone to overfitting and catastrophic forgetting, where task-specific updates degrade the model's original generative diversity. Addressing these challenges requires a delicate balance between adaptation flexibility and structural stability.

As illustrated in Fig. 2, current strategies for efficient adaptation generally fall into three paradigms, drawing inspiration from Vision Transformers (ViT) and Diffusion models. **(a) Adapter-based Methods** insert lightweight modules into the pre-trained backbone. Building on ViT adaptations [3, 19, 37], diffusion counterparts such as ControlNet [63, 65] and video-specific works like SimDA [60] introduce auxiliary networks for control signals or temporal features. However, these auxiliary modules introduce millions of trainable parameters, incurring substantial memory and computational overhead on large-scale models. **(b) Representation-based Methods** modify weight representations [18, 22] directly rather than adding modules. LoRA [14] employs low-rank decomposition, while RepAdapter [24] enables re-parameterization for inference efficiency. Yet, such **intrusive weight modifications** can disrupt pre-trained knowl-

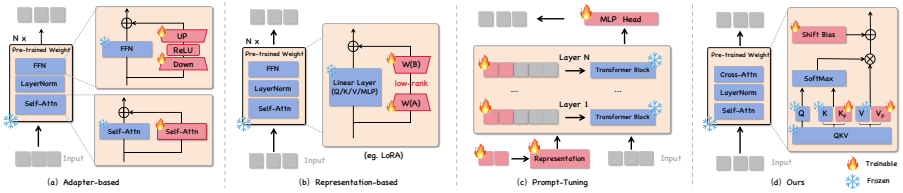

Fig. 2: Comparison of PEFT paradigms. (a) Adapter-based methods insert auxiliary modules in parallel alongside the backbone. (b) Representation-based methods (e.g., LoRA) approximate weight updates via low-rank decomposition. (c) Prompt-Tuning methods add trainable soft prompts to each layer and combine with the layer input. (d) Our method employs attention-embedded prompt tuning, injecting trainable prompts directly into attention mechanisms.

edge, leading to artifacts like inter-frame jitter. **(c) Prompt-Tuning Methods** [1, 16, 48, 53, 59, 64] guide attention with learnable prompts while freezing backbones, achieving high parameter efficiency in ViTs. However, their application to video diffusion models remains significantly under-explored, as the iterative denoising process poses unique challenges for maintaining temporal consistency and optimization stability. Unlike prior works, **(d) our approach** steers generation without modifying model weights, achieving comparable performance with parameter efficiency.

Despite this diversity, two critical limitations persist that hinder practical deployment. *(1) Huge model trainable parameters and high resource demands.* Existing PEFT methods still require training millions of parameters, which becomes prohibitive for models with more than 10 billion parameters. More critically, when scaling to massive video diffusion architectures, the cumulative memory footprint and computational overhead of these auxiliary modules remain substantial, rendering them impractical for practitioners with resource-constrained hardware. *(2) High Noise Incompatibility in condition-guided reward optimization.* While recent works like GRPOs [46, 61] apply reinforcement learning for alignment, they typically require full-model updates and group sampling, which are prohibitive for video diffusion. It also requires large labeled datasets and separate reward model training, which contradicts the goal of parameter and data efficiency. Furthermore, applying pixel-space rewards directly to latent representations is incompatible with high-noise denoising stages, causing optimization instability and a lack of timestep insensitivity. These gaps underscore the need for a more efficient framework that combines extreme parameter and data efficiency with stable optimization. Specifically, there is a pressing need for an *attention-embedded prompt tuning* mechanism that steers generation with minimal parameters to enhance attention performance without modifying pre-trained knowledge, alongside a *dual-space reward feedback optimization*, training via self-supervised latent objectives that enhance condition reward optimization. Addressing these challenges motivates our proposed MagicPrompt, which seeks

to balance adaptation flexibility with structural efficiency for high-quality video generation.

To bridge these gaps, we present MagicPrompt, a novel prompt-tuning framework engineered to achieve superior parameter efficiency and generalization capability for video diffusion models. Our core motivation is that strategic placement of lightweight soft prompts can effectively steer the model's attention mechanisms without modifying its core weights. This paradigm enables efficient adaptation by leveraging pre-trained knowledge for downstream tasks while helping mitigate the risk of catastrophic forgetting.

Architecturally, MagicPrompt operationalizes two design principles to systematically address the identified challenges: non-intrusive adaptation to preserve efficiency, and reward-guided optimization to promote alignment. To resolve parameter inefficiency and high resource demands, we eschew direct weight modification in favor of *Attention-Embedded Prompt Tuning*, injecting trainable soft prompts directly into attention mechanisms where feature integration occurs. In attention layers, learnable prompts prepended to Key and Value projections enable domain adaptation via attention redistribution, while frozen Query projections and backbone weights safeguard original generative capabilities. This restriction of updates to lightweight prompts reduces the number of trainable parameters by orders of magnitude compared to LoRA, enabling billion-scale adaptation on consumer hardware while inherently preventing forgetting. However, condition-video alignment requires specific reward models and is unstable during high-noise stage optimization. Thus, we integrate a *Dual-Space Reward Feedback Optimization* that supervises training via pixel-space perceptual rewards (HPS/MPS) for visual quality and latent-space objectives that enforce efficient denoising and significantly enhance generalization in condition-guided generation scenarios. Empirical validation confirms that MagicPrompt achieves competitive performance on video generation tasks while utilizing significantly fewer trainable parameters than LoRA and adapter-based methods. The contributions of this work are threefold:

- We propose **MagicPrompt**, a lightweight parameter-efficient tuning framework tailored for video diffusion models, enabling billion-scale model customization on resource-constrained hardware.
- We introduce *Attention-Embedded Prompt Tuning*, which injects trainable visual and textual prompts directly into the attention layers of video diffusion models. To overcome high noise incompatibility in reward training on the condition-guided generation task, we introduce a *Dual-Space Reward Feedback Optimization*. This non-intrusive approach preserves performance with orders of magnitude fewer parameters than existing methods and improves condition-guided reward training.
- Extensive experiments across text-to-video, image-to-video, and control-to-video tasks demonstrate that MagicPrompt achieves competitive performance with strong scalability across different model sizes.

2 Related Work

2.1 Video Diffusion Models

Diffusion models have revolutionized generative tasks by iteratively denoising random noise into structured data [11, 49]. Formally, the forward process diffuses data x_0 into noise x_T via a Markov chain, while the reverse process learns to recover x_0 by minimizing the noise prediction error $\mathcal{L} = \mathbb{E}_{x_0,t,\epsilon}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2]$. Extending this paradigm to video generation introduces significant complexity, as models must capture both spatial fidelity and temporal coherence within spatio-temporal tensors $V \in \mathbb{R}^{T\times H\times W\times C}$. Video diffusion models have evolved substantially since their inception, progressing from early 3D U-Net-based spatio-temporal modeling to the now-prevalent Transformer-centric architectural paradigm. Alongside architectural advances, the field has progressed in two complementary directions: reducing the computational cost of video synthesis and enhancing controllability for flexible conditional generation. Pioneering works established the 3D U-Net framework for video diffusion [6, 9, 12, 28, 31, 35, 40–42, 50–52, 55, 68, 68–70], and the scalable DiT paradigm later drove the shift to Transformer backbones, adopted by subsequent models such as Latte and CogVideo [26, 38, 62]. On the efficiency front, VideoLDM and VideoFusion proposed latent space optimization and multi-scale diffusion strategies, respectively [2, 25]; for controllable generation, Make-A-Video, AnimateDiff, and ControlVideo have successively advanced text-guided synthesis and explicit motion control [4, 7, 8, 10, 27, 47, 56, 57, 67]. However, as these models scale to billions of parameters to achieve higher fidelity, full fine-tuning becomes increasingly prohibitive, necessitating efficient adaptation strategies.

2.2 PEFT for Generative Models

Parameter-Efficient Fine-Tuning (PEFT) has become indispensable for adapting large generative models, addressing the prohibitive computational costs and storage requirements of full fine-tuning. In the realm of image generation, seminal works like *DreamBooth* [44] enable personalized synthesis by fine-tuning specific embeddings, while *DiffLoRA* [58] adapts low-rank decomposition for diffusion pipelines. These methods typically fall into three categories: **Representation-based** approaches, such as *LoRA* [14], approximate weight updates via low-rank matrices using $<1\%$ trainable parameters; **Adapter-based** methods [13] insert lightweight modules between layers for task-specific adaptation; and **Prompt-based** techniques like *Prefix Tuning* [21] learn continuous embeddings to steer generation without modifying model weights. Extending these paradigms to **video diffusion models** introduces unique challenges due to temporal dynamics and increased computational load. Recent adaptations include video-specific LoRA variants and adapter-based frameworks such as *SimDA* [60] and *VACE* [17], which incorporate modules to handle spatio-temporal features. While these methods reduce training costs by orders of magnitude and preserve pre-trained capabilities, they remain limited in few-shot scenarios. Specifically, ex-

isting video PEFT methods often struggle with the computational resource demands of generative diversity and text-video alignment when labeled data is scarce. This gap highlights the need for a more robust tuning framework that balances parameter efficiency with strong generalization capabilities in data-scarce regimes.

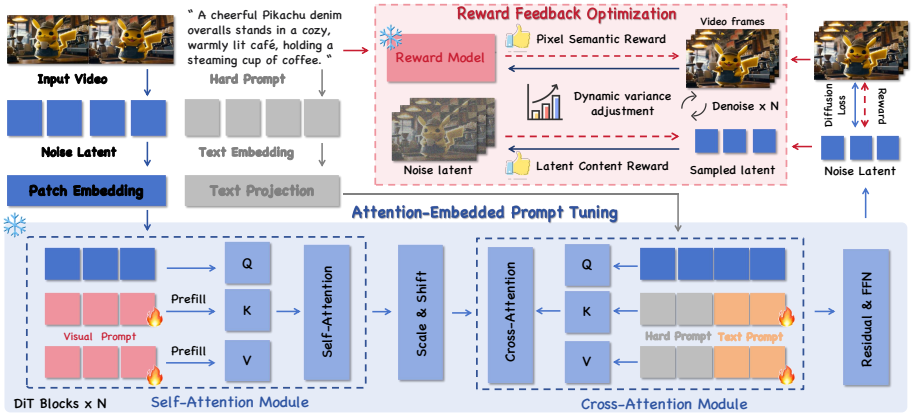

Fig. 3: Framework overview of MagicPrompt. Visual soft prompt prefill with latent feature in the key and value in the self-attention module, and textual soft prompt combined with text embedding in the cross-attention module to reweight the attention distribution. To better align with human preference, the reward feedback optimization is guided by both pixel-space semantic reward and latent objectives.

3 Method

Our method consists of two core components, as illustrated in Figure 3. In Section 3.1, we describe the definition of the problem. Section 3.2 introduces the Attention-Embedded Prompt Tuning method, which strategically inserts trainable soft prompts into both self-attention and cross-attention modules to efficiently adapt the pre-trained model to downstream tasks with minimal parameter overhead. Section 3.3 presents the Dual-Space Reward Feedback Optimization, which provides pixel-space rewards via HPS and MPS scores, and latent-space rewards through diffusion latent analysis to guide soft prompt learning toward high-quality generation.

3.1 Problem Definition

Let $\mathcal{M}_\theta$ denote a pre-trained video diffusion model with *frozen* parameters θ, generating video frames V conditioned on input C. Given target data $\mathcal{D}_{\text{target}}$, our goal is to adapt $\mathcal{M}_\theta$ to the target domain by learning *only* task-specific soft

prompts $\mathcal{P} = \{P_{txt}, P_{vis}, P_{bias}\}$. $P_{txt} \in \mathbb{R}^{L \times d}$ (textual soft prompts) are injected before text embeddings to guide cross-attention, and $P_{vis} \in \mathbb{R}^{L \times d}$ (visual soft prompts) are inserted into self-attention key-value sequences with shift biases term. Crucially, θ remains *untrained* while $\mathcal{P}$ is optimized to minimize:

$$\mathcal{L} = \mathbb{E}_{(V_0, C) \sim \mathcal{D}_{\text{target}}} \left[\|\epsilon - \epsilon_{\theta, \mathcal{P}}(V_t, t, C)\|^2 \right], \tag{1}$$

where $\epsilon_{\theta,\mathcal{P}}$ denotes the noise prediction with prompt-enhanced attention and V_t is the noisy latent/video representation obtained from V_0 at timestep t, and ϵ denotes the sampled Gaussian noise.

3.2 Attention-Embedded Prompt Tuning

Visual Prompt In video diffusion transformers, self-attention processes spatiotemporal tokens, where long-range temporal dependencies are essential for motion coherence. However, directly tuning backbone parameters incurs prohibitive computational costs while risking catastrophic forgetting of pre-trained dynamics. To resolve this, we inject task-specific visual prompts $P_{vis}^{\text{key}}, P_{vis}^{\text{value}} \in \mathbb{R}^{L_{vis} \times d}$ into the key-value sequences of each DiT block:

$$\begin{aligned} \tilde{\mathbf{K}}_{vis} &= [\mathbf{K} \oplus P_{vis}^{\text{key}}], \mathbf{K} \in \mathbb{R}^{B \times Seq \times d}, P_{vis}^{\text{key}} \in \mathbb{R}^{L_{vis} \times d}, \\ \tilde{\mathbf{V}}_{vis} &= [\mathbf{V} \oplus P_{vis}^{\text{value}}], \mathbf{V} \in \mathbb{R}^{B \times Seq \times d}, P_{vis}^{\text{value}} \in \mathbb{R}^{L_{vis} \times d}, \end{aligned} \tag{2}$$

where $\oplus$ denotes concatenation, $\tilde{\mathbf{K}}_{vis}$ and $\tilde{\mathbf{V}}_{vis}$ in attention layer apply to each DiT Block. This design preserves pre-trained spatial features while enabling dynamic attention redistribution across frames.

$$\mathbf{X} = \text{Self-Attention}(\mathbf{Q}, \tilde{\mathbf{K}_{vis}}, \tilde{\mathbf{V}_{vis}}) \tag{3}$$

Crucially, the visual prompts serve as temporal anchors: during denoising, queries from motion-critical regions (e.g., moving objects) develop stronger attention to P_{vis}, thereby explicitly guiding the model to prioritize relevant temporal patterns without modifying the frozen backbone. Consequently, the model achieves efficient adaptation with minimal parameter overhead ($L \cdot d \ll |\theta|$), directly addressing the resource constraints of billion-parameter video generators.

Shift Bias While visual prompts redistribute attention, they introduce latent space distribution shifts that manifest as frame inconsistency (e.g., jittery motions or abrupt scene transitions). This occurs because the concatenated prompts alter the statistical properties of the attention output, destabilizing the denoising trajectory across frames. To mitigate this, we introduce a learnable shift bias mechanism that calibrates the latent distribution through

$$\mathbf{X}' = \gamma \cdot \mathbf{X} + \beta, \quad \gamma \in \mathbb{R}^{1 \times 1 \times d}, \beta \in \mathbb{R}^{1 \times 1 \times d} \tag{4}$$

where γ and β are the layer-specific scale and bias trainable vectors. The scale γ amplifies weak but meaningful attention patterns (e.g., subtle object trajectories)

that would otherwise be suppressed by noise, while the bias β corrects systematic offsets in denoising trajectories (e.g., consistent motion direction errors). This dual calibration ensures that model weights remain bounded, thereby helping improve temporal coherence. Thus, the shift bias alleviates frame inconsistency introduced by visual prompts, forming a closed-loop optimization for stable video generation.

Textual Prompt In cross-attention, text embeddings $\mathbf{Context} \in \mathbb{R}^{L_c \times d}$ condition the diffusion process by aligning latent tokens with language semantics. However, unmodified embeddings often fail to precisely steer attention toward semantically critical regions, resulting in misaligned generations (e.g., the prompt "a dog running in a park" producing static dogs). To address this, we prepend task-aware textual prompts $P_{txt} \in \mathbb{R}^{L_{txt} \times d}$ to the text embeddings, forming $\tilde{\mathbf{Context}} = [P_{txt} \oplus \mathbf{Context}]$, where $L_{txt} \ll L_c$ denotes the prompt length. The augmented embeddings are then projected into key-value sequences for cross-attention:

$$\begin{aligned} \tilde{\mathbf{K}}_{txt} &= \tilde{\mathbf{Context}}\mathbf{W}_K, \\ \tilde{\mathbf{V}}_{txt} &= \tilde{\mathbf{Context}}\mathbf{W}_V, \end{aligned} \tag{5}$$

where $\mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ are frozen key/value projection matrices in the cross-attention layers of the pre-trained diffusion model. The cross-attention output is computed as:

$$\mathbf{X}_{cross} = \text{Softmax}\left(\frac{\mathbf{Q}\tilde{\mathbf{K}}_{txt}^{\top}}{\sqrt{d}}\right)\tilde{\mathbf{V}}_{txt} \tag{6}$$

During optimization, P_{txt} learns to function as a semantic filter: it suppresses attention to irrelevant tokens while amplifying attention to relevant tokens. This reshapes the attention distribution such that queries from motion-critical latent regions develop a stronger affinity toward contextually salient text tokens. Crucially, by prepending prompts rather than modifying the original embeddings, we preserve the integrity of pre-trained language representations while enabling task-specific adaptation. This design achieves robust text-video alignment with minimal parameter overhead ($L_{txt} \cdot d \ll |\theta|$), directly addressing semantic misalignment without altering the frozen text encoder.

3.3 Dual Reward Feedback Optimization

Pixel-Space Reward To directly optimize perceptual quality and semantic alignment, we compute rewards in pixel space by decoding denoised latents to video frames. Following the practice of DRaFT [5,39], we first sample key frames from the denoised video at low-noise timesteps ($t \geq 3T/4$, in inference mode), where visual semantics are sufficiently clear for reliable evaluation. These frames are then decoded via the VAE to obtain RGB frames $\hat{V} \in \mathbb{R}^{F \times H \times W \times 3}$. We employ two established perceptual metrics: the Human Preference Score (HPS)

measures text-image alignment by evaluating CLIP similarity between generated frames and the input text prompt, while the Motion Preference Score (MPS) assesses temporal coherence by computing optical flow consistency across consecutive frames. The pixel-space reward is defined as follows:

$$R_{\text{pixel}} = \alpha \cdot \text{HPS}(\hat{V}, C) + (1 - \alpha) \cdot \text{MPS}(\hat{V}), \tag{7}$$

where α balances semantic fidelity and motion quality. This reward provides direct supervision on visual quality but suffers from three limitations: first, computing rewards at early diffusion steps (high-noise stages) yields unreliable signals due to semantic ambiguity; second, reward values fluctuate significantly across timesteps, destabilizing training; third, deploying external reward models (e.g., HPS) incurs additional computational overhead. These limitations motivate our latent-space reward design to combine with the pixel reward.

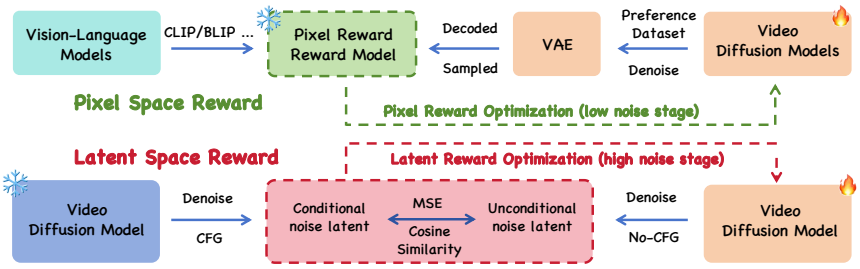

Fig. 4: Overview of dual reward feedback optimization. To enhance generalization in reward training, MagicPrompt integrates regularization signals from two domains. *Pixel Space:* Perceptual rewards (HPS, MPS) evaluate decoded frames to ensure visual fidelity. *Latent Space:* Complementary objectives enforce efficient denoising trajectories and robust conditioning via CFG.

Latent-Space Reward To overcome the pixel-space reward cannot be calculated during high-noise stages and the computational burden of pixel-space rewards, we propose a latent-space reward that leverages intrinsic properties of the diffusion process without requiring external reward models. Our design is grounded in two empirical observations: (1) Classifier-free guidance (CFG) consistently yields higher generation quality than unconditional sampling, indicating that CFG trajectories encode richer conditional semantics; (2) Latent representations become more coherent as noise level decreases, meaning the denoising process itself implicitly encodes quality progression along the timestep dimension. Based on these insights, we construct our latent reward by comparing two parallel denoising trajectories derived from the same input at the same timestep t: a CFG-guided trajectory and an unconditional trajectory. Specifically, we extract latent representations from both trajectories at timestep t, denoted z_t^{cfg}

and z_t^{uncond}, and compute their similarity as the reward signal:

$$R_{\text{latent}} = 1 - \text{MSE}\left(z_t^{\text{cfg}}, z_t^{\text{uncond}}\right), \tag{8}$$

This formulation follows two core design principles. First, the CFG trajectory at timestep t serves as the target quality level we aim for the model to achieve. Second, the unconditional trajectory at the same timestep acts as a baseline that represents unguided, noisier generation. Minimizing the distance between the two encourages the model to reach CFG-level quality even with weak guidance, effectively accelerating training convergence. Crucially, this reward is robust to noise level variations and timestep-agnostic: it operates on relative similarity between latents rather than absolute pixel values, so it remains effective across all denoising stages and eliminates the need for carefully designed reward scheduling. The final combined reward integrates both objectives:

$$R = \lambda_{\text{pixel}} R_{\text{pixel}} + \lambda_{\text{latent}} R_{\text{latent}}, \tag{9}$$

where λ_{pixel} and λ_{latent} balance pixel-level perceptual fidelity and latent-space semantic consistency. This dual reward mechanism stabilizes training while ensuring high-quality video generation.

4 Experiment

4.1 Experimental Setup

Setup. We evaluate our method on three video generation tasks. For text-to-video and image-to-video generation, we use the OpenVid dataset [36], which contains diverse video-text pairs covering various scenes and motions. We randomly split the dataset into training and test sets with a 9:1 ratio to ensure fair evaluation. For control-to-video generation, we train on the OpenHumanVid dataset [20], which provides human-centric videos with control signals (e.g., canny, depth), and evaluate on the TikTok dataset [15], a challenging benchmark featuring complex human motions and diverse backgrounds. Our pre-trained weights are from Wan2.1-Fun-InP and Wan2.2-TI2V-5B [55]. For fair comparison, we standardize the training configuration across all methods. **LoRA** is configured with Rank=64, α=32, and learning rates of 1e-4 (1.3B/5B) and 3e-5 (14B). **MagicPrompt** uses 64 soft prompt tokens with learning rates of 5e-3 (1.3B/5B) and 1e-3 (14B).

Evaluation Metrics. We employ a comprehensive set of four metrics to assess generation quality from multiple perspectives. For semantic alignment, we compute the CLIP Score [43] between generated frames and input conditions (text prompts or control signals), where higher values indicate better text-video or control-video correspondence. For perceptual quality, we adopt LPIPS [66] to measure the perceptual similarity between generated and reference frames, with lower scores reflecting higher visual fidelity. For distribution-level realism, we

Method	T2V				I2V				Control2V			
	CLIP ↑	LPIPS ↓	FID ↓	FVD ↓	CLIP ↑	LPIPS ↓	FID ↓	FVD ↓	CLIP ↑	LPIPS ↓	FID ↓	FVD ↓
Pre-Trained	0.66	0.69	103.53	824.53	0.89	0.38	57.72	299.81	0.85	0.40	30.35	488.72
LoRA	0.68	**0.66**	101.76	853.56	0.91	**0.35**	43.19	**269.02**	0.86	0.39	**26.09**	419.82
VACE	0.65	0.71	99.64	760.33	0.88	0.44	43.04	354.67	0.86	**0.41**	39.29	464.45
Ours	**0.68**	0.66	**86.56**	**637.46**	**0.92**	0.37	**29.23**	274.88	**0.87**	0.39	34.43	**412.82**

Table 1: Quantitative comparison with baseline methods across three tasks: Text-to-Video (T2V), Image-to-Video (I2V), and Control-to-Video (Control2V). T2V and I2V use the 1.3B model, whereas Control2V adopts 14B model.

use Fréchet Inception Distance (FID) [45] to compare feature distributions of generated and real videos, where lower values denote better image quality. Finally, for temporal coherence, we report Fréchet Video Distance (FVD) [54] to assess motion smoothness and video-level consistency, with lower scores indicating superior temporal dynamics. All metrics are computed on held-out test sets, with higher CLIP Score and lower LPIPS/FID/FVD indicating better overall performance.

4.2 Quantitative Experiments

We conduct quantitative evaluations across T2V, I2V, and Control2V test sets, comparing MagicPrompt against three representative baselines. As shown in Table 1, MagicPrompt achieves competitive or superior performance across all tasks while updating the fewest parameters.

Analysis of Comparison. As shown in Tab. 1, in Text-to-Video generation, our method matches LoRA in CLIP Score and LPIPS, while substantially outperforming all baselines in distribution quality, achieving the best FID and FVD. This indicates that our attention-embedded prompts effectively adapt semantic alignment without compromising visual fidelity. In Image-to-Video, MagicPrompt attains the highest CLIP Score and the lowest FID, demonstrating strong preservation of input image semantics. While LoRA shows a marginal advantage in FVD, our method remains highly competitive with significantly fewer trainable parameters. For Control-to-Video, MagicPrompt achieves the best FVD, outperforming VACE and LoRA while maintaining comparable CLIP and LPIPS scores, highlighting its robustness in handling complex control signals. Critically, these results are achieved with orders of magnitude fewer trainable parameters than LoRA or VACE. This efficiency-performance trade-off makes MagicPrompt particularly suitable for resource-constrained adaptation of billion-scale video diffusion models.

Analysis of Reward Feedback. To isolate the impact of reward strategies, we maintain the trainable parameters fixed as visual and textual soft prompts throughout this experiment. As detailed in Table 2, our Dual-Space Reward

Method	CLIP ↑	LPIPS ↓	SSIM ↑	PSNR ↑	FID ↓	FVD ↓
Latent Reward	0.92	0.27	**0.81**	19.57	21.46	217.91
Pixel Reward	0.89	0.29	0.77	18.90	18.81	142.06
Dual Reward	**0.93**	**0.27**	0.79	**19.64**	**17.92**	**141.44**
No Reward	0.92	0.37	0.71	18.21	29.12	274.88

Table 2: Comparison of different reward strategies and training configurations. Our full method achieves competitive performance compared to Pixel Reward and Latent Reward. ↑ indicates higher is better, ↓ indicates lower is better.

Feedback Optimization enables the model to maintain robust condition-video alignment. Our ablation study demonstrates that pixel-space and latent-space rewards each possess distinct and complementary strengths. When integrated into the Dual Reward framework, their combined effect achieves the best CLIP, PSNR, FID, and FVD, while remaining competitive in LPIPS and SSIM. This demonstrates that latent-space objectives effectively regularize the early denoising trajectory, allowing for fewer sampling steps without collapsing generation quality. Furthermore, compared to No Reward (without reward training), our method greatly enhances the generation effect despite utilizing significantly limited trainable parameters. These results suggest that under resource-constrained scenarios, combining minimal parameter updates with reward feedback offers a viable paradigm for achieving efficient training.

		Image2Video Task				**Text2Video Task**			
Model Size	**Method**	**CLIP (↑)**	**LPIPS (↓)**	**FID (↓)**	**FVD (↓)**	**CLIP (↑)**	**LPIPS (↓)**	**FID (↓)**	**FVD (↓)**
Wan2.1-1.3B	Fine-Tuning	0.907	0.380	37.719	284.168	0.660	**0.654**	103.534	824.528
	LoRA	0.918	**0.354**	43.196	**269.021**	0.680	0.663	101.766	853.554
	Ours	**0.923**	0.371	**29.236**	274.880	**0.684**	0.661	**86.560**	**637.466**
Wan2.2-5B	Fine-Tuning	**0.912**	0.380	**29.019**	328.852	0.678	0.689	104.840	724.924
	LoRA	0.911	0.378	30.378	289.879	0.699	0.665	**90.734**	**662.144**
	Ours	0.910	**0.366**	29.797	**284.780**	**0.699**	**0.663**	94.241	694.937
Wan2.1-14B	Pre-trained	0.924	0.380	29.411	247.651	0.670	0.687	101.523	778.461
	LoRA	**0.933**	**0.351**	**25.781**	256.390	0.669	0.674	107.738	759.532
	Ours	0.930	0.356	28.378	**240.456**	**0.689**	**0.668**	**94.152**	**682.851**

Table 3: Performance comparison of different model sizes and methods on Image2Video and Text2Video tasks. ↑ indicates higher is better, and ↓ indicates lower is better.

Cross-Scale Scalability Evaluation. To further evaluate the scalability and generalization ability of MagicPrompt, we conduct experiments across three model sizes, namely 1.3B, 5B, and 14B parameters, on both Image-to-Video and Text-to-Video tasks. This evaluation aims to determine whether the proposed prompt-based adaptation remains effective as the capacity of the base model increases. As shown in Table 3, MagicPrompt achieves consistently competitive perfor-

Fig. 5: New condition adaptation. The pre-trained model suffers from notable performance degradation on unseen new conditions, as it has not learned to model the novel condition prior. After tuning with our method, the model achieves substantial performance gains and effectively adapts to the target condition.

mance across 1.3B, 5B, and 14B models on both I2V and T2V tasks. On larger backbones, especially 5B and 14B, our method improves FVD while using far fewer trainable parameters, demonstrating strong temporal quality with minimal parameter updates. In T2V, MagicPrompt shows particularly clear advantages on the 14B model, achieving the best results across all metrics, while LoRA does not consistently improve over the pre-trained baseline. The results show that MagicPrompt consistently achieves competitive performance across different model scales while requiring significantly fewer trainable parameters than conventional parameter-efficient fine-tuning methods.

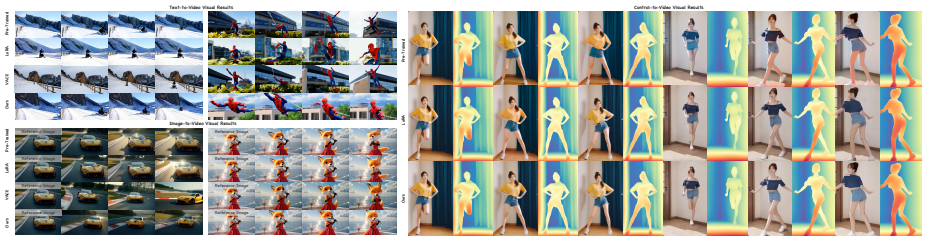

Fig. 6: Qualitative comparison with baselines. Our method demonstrates superior temporal coherence and visual fidelity across T2V, I2V, and Control2V tasks. Unlike baselines exhibiting motion blur and structural distortion, our method preserves fine details and ensures consistent motion dynamics.

4.3 Qualitative Experiments

Text-to-Video (T2V). As shown in Fig. 6, our method demonstrates superior adherence to textual prompts. In the sled scenario, baseline methods such as LoRA and VACE struggle to align motion semantics with the text, leading to inconsistent character movements. In the Spiderman scenario, competing methods exhibit severe motion blur and character distortion, whereas MagicPrompt

maintains sharp details and consistent identity throughout the sequence. This confirms that our attention-embedded prompts effectively guide semantic generation without disrupting spatial structure.

Image-to-Video (I2V). MagicPrompt excels in motion coherence and content preservation. In the racing car scenario, our method generates realistic, rapid motion dynamics, while baselines produce overly static or slow-moving results. In the anime scenario, MagicPrompt better preserves reference image details, such as the object held in the character's left hand, which is often lost or distorted by other methods. As illustrated in Fig. 7, our method adaptively adjusts the denoising distribution and consistently achieves better generation quality under a few-shot scenario.

Control-to-Video (Control2V). For condition-guided generation, our method achieves robust motion consistency. As shown in Fig. 5 and Fig. 6, MagicPrompt not only accurately follows the pose sequence but also adapts to new conditions without introducing structural distortion, unlike baselines that suffer from significant color blur and temporal jitter. These visual results corroborate our quantitative findings, demonstrating that MagicPrompt delivers high-quality, temporally coherent videos across diverse adaptation scenarios.

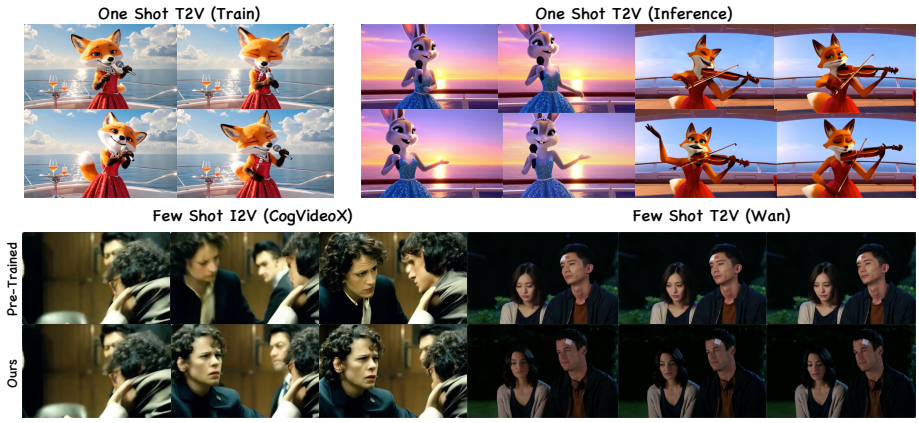

Fig. 7: Few-Shot result visualization. Under one-shot and few-shot training regimes, our method adaptively adjusts the denoising distribution and consistently achieves better generation quality.

4.4 Ablation Study

To validate the design choices of MagicPrompt, we conduct ablation studies on two critical aspects: (1) the contribution of core components, and (2) the

Component Ablation					Soft Token Number				
Method	CLIP ↑	LPIPS ↓	FID ↓	FVD ↓	Method	CLIP ↑	LPIPS ↓	FID ↓	FVD ↓
w/o soft prompt	0.91	0.37	30.94	284.16	4 token	0.91	0.39	32.76	347.43
w/o shift bias	0.89	0.39	36.23	365.64	8 token	0.91	0.40	34.73	351.18
Full(64 token)	**0.92**	**0.37**	**29.23**	**274.88**	32 token	0.90	0.40	33.88	353.72
Fine-Tuning	0.90	0.38	37.71	284.168	64 token	**0.92**	**0.37**	**29.23**	**274.88**

Table 4: Ablation study on components and soft token numbers.

sensitivity to the number of soft tokens. All experiments are performed on the Image-to-Video (I2V) task with the Wan2.1-1.3B model.

Component Analysis. As shown in the left panel of Table 4, removing either core component leads to performance degradation, confirming their complementary roles. Specifically, *w/o soft prompt* causes moderate declines in FID and FVD, indicating that prompt injection is essential for fine-grained semantic adaptation. More critically, *w/o shift bias* results in substantial drops across all metrics, particularly FID and FVD. This demonstrates that the shift bias plays a crucial role in stabilizing latent distribution alignment. The full model, combining both components with 64 soft tokens, achieves the best overall performance, validating our design principle of non-intrusive adaptation with attention distribution regularization.

Soft Token Number Sensitivity. The right panel of Table 4 examines the impact of soft token count on performance-efficiency trade-offs. With only 4 tokens, the model lacks sufficient capacity to capture task-specific features, resulting in suboptimal FID and FVD. Increasing the token number from 4 to 8 or 32 does not yield monotonic improvements, while 64 tokens provide a clear performance gain. Notably, 64 tokens achieve the best results across all metrics, suggesting that this capacity is sufficient to encode rich adaptation signals without overfitting. Importantly, even with 64 tokens, our trainable parameters remain orders of magnitude fewer than LoRA or Adapter-based methods, maintaining the core advantage of parameter efficiency.

5 Conclusion

We introduced MagicPrompt, an ultra-lightweight framework for adapting large-parameter video diffusion models with minimal trainable parameters ($< 1\%$). Experiments across diverse video tasks demonstrate that MagicPrompt achieves competitive visual results, compared to existing parameter-efficient methods, while reducing trainable parameters by orders of magnitude.

6 Acknowledgments

This work was supported by a grant from the NSFC/RGC Collaborative Research Scheme Project (No. CRS_HKUST605/25).

References

1. Bandara, W.G.C., Patel, V.M.: Attention Prompt Tuning: Parameter-efficient Adaptation of Pre-trained Models for Spatiotemporal Modeling (2024), https://arxiv.org/abs/2403.06978 3
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your Latents: High-resolution Video Synthesis with Latent Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023) 5
3. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting Vision Transformers for Scalable Visual Recognition. Advances in Neural Information Processing Systems **35**, 16664–16678 (2022) 2
4. Chen, Y., He, X., Ma, X., Ma, Y.: ContextFlow: Training-Free Video Object Editing via Adaptive Context Enrichment. arXiv preprint arXiv:2509.17818 (2025) 5
5. Clark, K., Vicol, P., Swersky, K., Fleet, D.: Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In: International Conference on Learning Representations. vol. 2024, pp. 4793–4822 (2024) 8
6. Deng, T., Chen, X., Chen, Y., Chen, Q., Xu, Y., Yang, L., Xu, L., Zhang, Y., Zhang, B., Huang, W., Wang, H.: GaussianDWM: 3D Gaussian Driving World Model for Unified Scene Understanding and Multi-Modal Generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10656–10667 (June 2026) 5
7. Deng, T., Chen, Y., Zhang, L., Yang, J., Yuan, S., Liu, J., Wang, D., Wang, H., Chen, W.: Compact 3d gaussian splatting for Dense Visual SLAM. arXiv preprint arXiv:2403.11247 (2024) 5
8. Feng, K., Ma, Y., Wang, B., Qi, C., Chen, H., Chen, Q., Wang, Z.: Dit4edit: Diffusion Transformer for Image Editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2969–2977 (2025) 5
9. Gao, H., Tang, B., Song, Y., Fang, G., He, Z., Yang, J., Shou, M.Z.: PAI-Studio: Cinematic Video Background Replacement with Camera-Aware Motion. arXiv preprint arXiv:2606.01399 (2026) 5
10. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning (2024), https://arxiv.org/abs/2307.04725 5
11. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. Advances in neural information processing systems **33**, 6840–6851 (2020) 5
12. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video Diffusion Models (2022), https://arxiv.org/abs/2204.03458 5
13. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-Efficient Transfer Learning for NLP. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019) 5
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-Rank Adaptation of Large Language Models. Iclr **1**(2), 3 (2022) 2, 5

15. Jafarian, Y., Park, H.S.: Learning High Fidelity Depths of Dressed Humans by Watching Social Media Dance Videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12753–12762 (June 2021) 10
16. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual Prompt Tuning (2022), https://arxiv.org/abs/2203.12119 3
17. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: VACE: All-in-one Video Creation and Editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025) 5
18. Jie, S., Deng, Z.H.: Fact: Factor-tuning for Lightweight Adaptation on Vision Transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1060–1068 (2023) 2
19. Jie, S., Deng, Z.H., Chen, S., Jin, Z.: Convolutional Bypasses are Better Vision Transformer Adapters. In: ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain–Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024). pp. 202–209. SAGE Publications Pvt. Ltd 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP (2024) 2
20. Li, H., Xu, M., Zhan, Y., Mu, S., Li, J., Cheng, K., Chen, Y., Chen, T., Ye, M., Wang, J., et al.: Openhumanvid: A Large-Scale High-Quality Dataset for Enhancing Human-Centric Video Generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7752–7762 (2025) 10
21. Li, X.L., Liang, P.: Prefix-tuning: Optimizing Continuous Prompts for Generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4582–4597 (2021) 5
22. Lian, D., Zhou, D., Feng, J., Wang, X.: Scaling & Shifting Your Features: A New Baseline for Efficient Model Tuning. Advances in Neural Information Processing Systems **35**, 109–123 (2022) 2
23. Liu, J., Wang, X., Lin, Y., Wang, Z., Wang, P., Cai, P., Zhou, Q., Yan, Z., Yan, Z., Shi, Z., et al.: A Survey on Cache Methods in Diffusion Models: Toward Efficient Multi-Modal Generation. arXiv preprint arXiv:2510.19755 (2025) 2
24. Luo, G., Huang, M., Zhou, Y., Sun, X., Jiang, G., Wang, Z., Ji, R.: Towards Efficient Visual Adaption via Structural Re-parameterization (2023), https://arxiv.org/abs/2302.08106 2
25. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation (2023), https://arxiv.org/abs/2303.08320 5
26. Ma, X., Wang, Y., Chen, X., Jia, G., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent Diffusion Transformer for Video Generation (2025), https://arxiv.org/abs/2401.03048 5
27. Ma, Y., Cun, X., Liang, S., Xing, J., He, Y., Qi, C., Chen, S., Chen, Q.: Magicstick: Controllable Video Editing via Control Handle Transformations. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 9385–9395. IEEE (2025) 5
28. Ma, Y., Feng, K., Hu, Z., Wang, X., Wang, Y., Zheng, M., He, X., Zhu, C., Liu, H., He, Y., et al.: Controllable Video Generation: A Survey. arXiv preprint arXiv:2507.16869 (2025) 5
29. Ma, Y., Feng, K., Zhang, X., Liu, H., Zhang, D.J., Xing, J., Zhang, Y., Yang, A., Wang, Z., Chen, Q.: Follow-Your-Creation: Empowering 4D Creation through Video Inpainting. arXiv preprint arXiv:2506.04590 (2025) 2

30. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow Your Pose: Pose-guided Text-to-video Generation Using Pose-free Videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024) 2
31. Ma, Y., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., Liu, W., et al.: Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–12 (2024) 5
32. Ma, Y., Liu, Y., Zhu, Q., Yang, A., Feng, K., Zhang, X., Li, Z., Han, S., Qi, C., Chen, Q.: Follow-Your-Motion: Video Motion Transfer via Efficient Spatial-Temporal Decoupled Finetuning. arXiv preprint arXiv:2506.05207 (2025) 2
33. Ma, Y., Wang, X., Ma, Q., Wang, Q., Zheng, M., Yang, X., Li, H., Zhao, C., Ying, J., Yang, H., et al.: Group Editing: Edit Multiple Images in One Go. arXiv preprint arXiv:2603.22883 (2026) 2
34. Ma, Y., Wang, Z., Ren, T., Zheng, M., Liu, H., Guo, J., Fong, M., Xue, Y., Zhao, Z., Schindler, K., et al.: FastVMT: Eliminating Redundancy in Video Motion Transfer. arXiv preprint arXiv:2602.05551 (2026) 2
35. Ma, Y., Yan, Z., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., et al.: Follow-your-emoji-faster: Towards Efficient, Fine-Controllable, and Expressive Freestyle Portrait Animation. arXiv preprint arXiv:2509.16630 (2025) 5
36. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: OpenVid-1M: A Large-Scale High-quality Dataset for Text-to-video Generation. In: International Conference on Learning Representations. vol. 2025, pp. 1045–1064 (2025) 10
37. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: ST-Adapter: Parameter-Efficient Image-to-Video Transfer Learning. Advances in Neural Information Processing Systems **35**, 26462–26477 (2022) 2
38. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023) 5
39. Prabhudesai, M., Goyal, A., Pathak, D., Fragkiadaki, K.: Aligning Text-to-Image Diffusion Models with Reward Backpropagation (2024), https://arxiv.org/abs/2310.03739 8
40. Qiu, X., Hu, J., Zhou, L., Wu, X., Du, J., Zhang, B., Guo, C., Zhou, A., Jensen, C.S., Sheng, Z., Yang, B.: TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. In: Proc. VLDB Endow. pp. 2363–2377 (2024) 5
41. Qiu, X., Wu, X., Cheng, H., Liu, X., Guo, C., Hu, J., Yang, B.: DBLoss: Decomposition-based Loss Function for Time Series Forecasting. In: NeurIPS (2025) 5
42. Qiu, X., Wu, X., Lin, Y., Guo, C., Hu, J., Yang, B.: DUET: Dual clustering enhanced multivariate time series forecasting. In: SIGKDD. pp. 1185–1196 (2025) 5
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable Visual Models from Natural Language Supervision. In: International Conference on Machine Learning. pp. 8748–8763. PmLR (2021) 10
44. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023) 5

45. Seitzer, M.: pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid (August 2020), version 0.3.0 11
46. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y., Wu, Y., Guo, D.: DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models (2024), https://arxiv.org/abs/2402.03300 3
47. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y.: Make-A-Video: Text-to-Video Generation without Text-Video Data (2022), https://arxiv.org/abs/2209.14792 5
48. Sohn, K., Chang, H., Lezama, J., Polania, L., Zhang, H., Hao, Y., Essa, I., Jiang, L.: Visual Prompt Tuning for Generative Transfer Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19840–19851 (2023) 3
49. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models (2022), https://arxiv.org/abs/2010.02502 5
50. Song, Y., Huang, S., Yao, C., Ci, H., Ye, X., Liu, J., Zhang, Y., Shou, M.Z.: ProcessPainter: Learning to Draw from Sequence Data. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–10 (2024) 5
51. Song, Y., Liu, C., Jiang, Y., Shou, M.Z.: StreamingEffect: Real-Time Human-Centric Video Effect Generation. arXiv preprint arXiv:2605.17019 (2026) 5
52. Song, Y., Yao, W., Wang, H., Shou, M.Z.: VISTA: Triplet-Supervised Video Style Transfer with Diffusion Transformers. arXiv preprint arXiv:2605.17312 (2026) 5
53. Sun, G., Mendieta, M., Luo, J., Wu, S., Chen, C.: Fedperfix: Towards Partial Model Personalization of Vision Transformers in Federated Learning. In: ConferenceProceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4988–4998 (2023) 3
54. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards Accurate Generative Models of Video: A New Metric & Challenges (2019), https://arxiv.org/abs/1812.01717 11
55. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and Advanced Large-Scale Video Generative Models (2025), https://arxiv.org/abs/2503.20314 5, 10
56. Wang, J., Ma, Y., Guo, J., Xiao, Y., Huang, G., Li, X.: Cove: Unleashing the Diffusion Feature Correspondence for Consistent Video Editing. Advances in Neural Information Processing Systems **37**, 96541–96565 (2024) 5
57. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., Chen, Y., Li, X., Shan, Y.: Taming Rectified Flow for Inversion and Editing. arXiv preprint arXiv:2411.04746 (2024) 5
58. Wu, Y., Shi, Y., Wei, J., Sun, C., Yang, Y., Shen, H.T.: DiffLoRA: Generating Personalized Low-Rank Adaptation Weights with Diffusion (2024), https://arxiv.org/abs/2408.06740 5
59. Xia, J., Zhang, C., Zhang, Y., Zhou, C., Wang, Z., Liu, B., Yin, D.: DAPE: Dual-Stage Parameter-Efficient Fine-Tuning for Consistent Video Editing with Diffusion Models (2025), https://arxiv.org/abs/2505.07057 3

60. Xing, Z., Dai, Q., Hu, H., Wu, Z., Jiang, Y.G.: SimDA: Simple Diffusion Adapter for Efficient Video Generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7827–7839 (2024) 2, 5
61. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., Luo, P.: DanceGRPO: Unleashing GRPO on Visual Generation (2025), https://arxiv.org/abs/2505.07818 3
62. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: CogvideoX: Text-to-Video Diffusion Models with an Expert Transformer. In: International Conference on Learning Representations. vol. 2025, pp. 83048–83077 (2025) 5
63. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models (2023), https://arxiv.org/abs/2308.06721 2
64. Yu, B.X.B., Chang, J., Liu, L., Tian, Q., Chen, C.W.: Towards a Unified View on Visual Parameter-Efficient Transfer Learning (2023), https://arxiv.org/abs/2210.00788 3
65. Zhang, L., Rao, A., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 2
66. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric (2018), https://arxiv.org/abs/1801.03924 10
67. Zhang, Y., Wei, Y., ZHANG, X., Zuo, W., Tian, Q., et al.: ControlVideo: Training-free Controllable Text-to-Video Generation. In: International Conference on Learning Representations. vol. 2024, pp. 54441–54461 (2024) 5
68. Zhang, Y., Ma, Y., Wang, B., Chen, Q., Wang, Z.: MagicColor: Multi-Instance Sketch Colorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15205–15217 (2025) 5
69. Zhang, Y., Ma, Y., Wang, B., Feng, K., Jin, Y., Chen, Q., Rao, A., Wang, Z.: InstanceAnimator: Multi-Instance Sketch Video Colorization. arXiv preprint arXiv:2603.25357 (2026) 5
70. Zhang, Y., Ma, Y., Yi, F., Qi, C., Zhang, C., Feng, K., Wang, Z.: Tea-Adapter: Teacher Adapter for Efficient Conditional Generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4805–4815 (2026) 5
71. Zheng, S., Feng, L., Wang, X., Zhou, Q., Cai, P., Zou, C., Liu, J., Lin, Y., Chen, J., Ma, Y., et al.: Forecast then Calibrate: Feature Caching as ODE for Efficient Diffusion Transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13449–13457 (2026) 2