I4VGEN: IMAGE AS FREE STEPPING STONE FOR TEXT-TO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

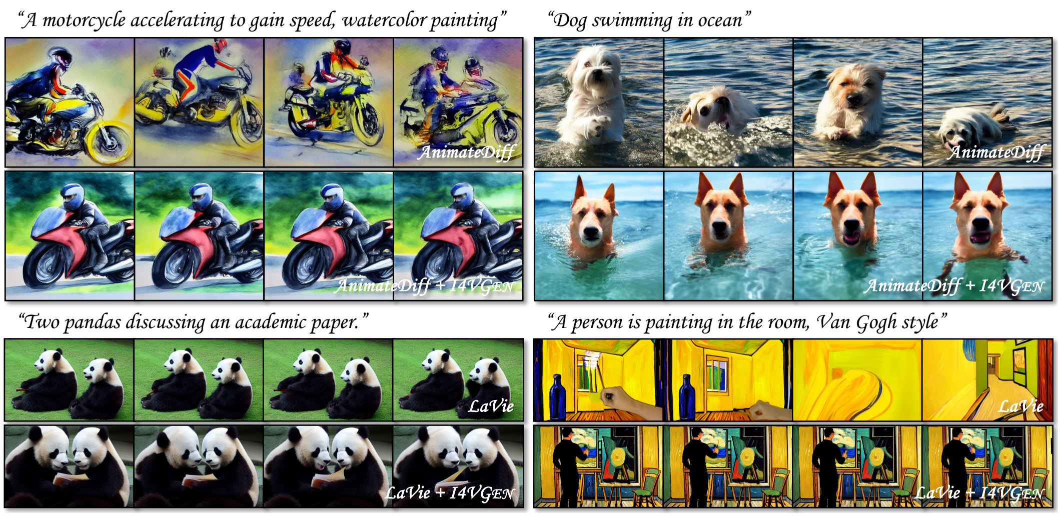


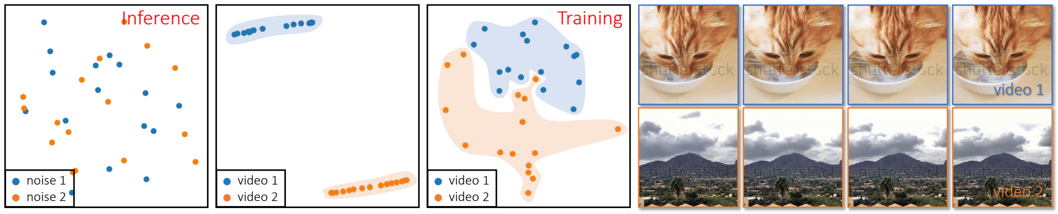
Figure 1: **Example results** synthesized by the proposed I4VGEN. I4VGEN is seamlessly integrated into existing pre-trained text-to-video diffusion models without additional training, significantly improving the temporal consistency (*e.g.*, top-left and bottom-right), visual realism (*e.g.*, top-right), and semantic fidelity (*e.g.*, bottom-left) of the synthesized videos.

ABSTRACT

Text-to-video generation has trailed behind text-to-image generation in terms of quality and diversity, primarily due to the inherent complexities of spatio-temporal modeling and the limited availability of video-text datasets. Recent text-to-video diffusion models employ the image as an intermediate step, significantly enhancing overall performance but incurring high training costs. In this paper, we present I4VGEN, a novel video diffusion inference pipeline to leverage advanced image techniques to enhance pre-trained text-to-video diffusion models, which requires no additional training. Instead of the vanilla text-to-video inference pipeline, I4VGEN consists of two stages: anchor image synthesis and anchor image-augmented text-to-video synthesis. Correspondingly, a simple yet effective generation-selection strategy is employed to achieve visually-realistic and semantically-faithful anchor image, and an innovative noise-invariant video score distillation sampling (NI-VSDS) is developed to animate the image to a dynamic video by distilling motion knowledge from video diffusion models, followed by a video regeneration process to refine the video. Extensive experiments show that the proposed method produces videos with higher visual realism and textual fidelity. Furthermore, I4VGEN also supports being seamlessly integrated into existing image-to-video diffusion models, thereby improving overall video quality.

1 INTRODUCTION

Recent advances in large-scale text-to-image diffusion models (Esser et al., 2021; Balaji et al., 2022; Ramesh et al., 2022; Nichol et al., 2022; Saharia et al., 2022; Feng et al., 2023; Gu et al., 2023; Xue



(a) pure Gaussian noise (b) real video (c) noisy video (d) sampled videos

Figure 2: **Illustration of non-zero terminal signal-to-noise ratio.** We employ t-SNE to visualize the distributions of pure Gaussian noise, real video, and noisy video at the timestep T, where each data point represents an independently sampled noise point or video frame. The noise schedule of AnimateDiff (Guo et al., 2024b) is used, and all operations are performed in the latent space of the video autoencoder. (a) The distribution of pure Gaussian noise exhibits a disordered and diffuse nature; (b) real videos are temporally-correlated and different videos can be clearly distinguished from each other; (c) noisy videos preserve a certain degree of temporal correlation and maintain separability between different videos; (d) sampled videos for visualization.

et al., 2023) have demonstrated the capability to generate diverse and high-quality images from extensive web-scale image-text pair datasets. Efforts to extend these diffusion models to text-to-video synthesis (Ho et al., 2022a; Zhou et al., 2022; Chen et al., 2023a; Singer et al., 2023; Wang et al., 2023b;e; Blattmann et al., 2023a; Girdhar et al., 2023; Guo et al., 2024b; Bao et al., 2024) have involved leveraging video-text pairs and temporal modeling. However, text-to-video generation remains inferior to image counterpart in terms of both quality and diversity, primarily due to the complex nature of spatio-temporal modeling and the limited size of video-text datasets, which are often an order of magnitude smaller than image-text datasets.

This paper explores a novel video diffusion inference pipeline that leverages advanced image techniques to enhance pre-trained text-to-video diffusion models, focusing on the following two insights:

Image conditioning for text-to-video generation. Recent methods (Blattmann et al., 2023a; Zhang et al., 2023b; Girdhar et al., 2023; Chen et al., 2024a; Li et al., 2023; Hu et al., 2023) have adopted image-guided text-to-video generation, where an initial image generation step significantly enhances video output quality. This paradigm benefits from the strong capabilities of text-to-image models by using the generated images as detailed references for video synthesis. While effective, these approaches incurs additional high training costs. This paper builds on this insight but innovates by designing a novel video diffusion inference pipeline to leverage image information, thereby enhancing text-to-video generation performance without additional training expense.

Zero terminal-SNR noise schedule. A prevalent issue in diffusion models is the non-zero terminal signal-to-noise ratio (SNR) (Guttenberg; Lin et al., 2024). The mismatch between the training phase, where residual signals persist in noisy videos at the terminal diffusion timestep T, and the inference phase, which uses pure Gaussian noise at the timestep T, creates a gap that degrade the model performance. As illustrated in Fig. 2, noisy videos exhibit temporal correlation that is distinctly different from the independent and identically distributed pure Gaussian noise. This paper is dedicated to reconfiguring the inference pipeline to circumvent this issue.

Motivated by these insights, we propose a novel video diffusion inference pipeline, called I4VGEN, which enhances pre-trained text-to-video diffusion models by incorporating image information into the inference process. This method requires no additional learnable parameters and training costs, and can be seamlessly integrated into existing text-to-video diffusion models, circumventing the non-zero terminal SNR issue and improving output quality.

Specifically, instead of the vanilla text-to-video inference pipeline, which fails to leverage image reference information, I4VGEN decomposes the inference process into two stages: anchor image synthesis and anchor image-augmented text-to-video synthesis. For the former, a simple yet effective generation-selection strategy is introduced, which involves synthesizing candidate images and selecting the most suitable one using a reward-based mechanism, thereby obtaining a visually-realistic anchor image that is closely aligned with the text prompt. For the latter, we develop an innovative noise-invariant video score distillation sampling (NI-VSDS) to animate the anchor image to a dynamic video by extracting motion knowledge from text-to-video diffusion models, followed by a video regeneration process, *i.e.*, diffusion-denoising, to refine the video. This inference pipeline avoids the issue of non-zero terminal SNR.

Extensive quantitative and qualitative analyses demonstrate that I4VGEN can be effectively applied to various text-to-video diffusion models, significantly improving the temporal consistency, visual realism, and semantic fidelity of the synthesized videos (see Fig. 1). Moreover, our method can also be seamlessly integrated into existing image-to-video diffusion models, thereby enhancing the temporal consistency and visual quality of the generated videos (see Fig. 6).

The main novelties and contributions are as follows:

- We propose a novel video diffusion inference pipeline, called I4VGEN, which enhances pre-trained text-to-video diffusion models by incorporating image reference information into the inference process, without requiring additional training or learnable parameters.
- We employ a simple yet effective generation-selection strategy to achieve high-quality image, and design a novel noise-invariant video score distillation sampling for image animation.
- We comprehensively evaluate our approach with representative text-to-video diffusion models, and demonstrate I4VGEN significantly improves the quality of generated videos. Furthermore, I4VGEN can also be adapted to image-to-video diffusion models, leading to improved results.

2 PRELIMINARIES

Video diffusion models. Aligned with the framework of image diffusion models, Video diffusion models (VDMs) predominantly utilize the paradigm of latent diffusion models (LDMs). Unlike traditional methods that operate directly in the pixel space, VDMs function within the latent space defined by a video autoencoder. Specifically, a video encoder $\mathcal{E}\left(\cdot\right)$ learns the mapping from an input video $\mathbf{v} \in \mathcal{V}, \mathbf{v} = \{\mathbf{f}^1, \mathbf{f}^2, \cdots, \mathbf{f}^F\}$ to a latent code $\mathbf{z} = \mathcal{E}\left(\mathbf{v}\right) = \{\mathbf{z}^1, \mathbf{z}^2, \cdots, \mathbf{z}^f\}$. Subsequently, a video decoder $\mathcal{D}(\cdot)$ reconstructs the input video, aiming for $\mathcal{D}\left(\mathcal{E}\left(\mathbf{v}\right)\right) \approx \mathbf{v}$. Typically, image autoencoder is used in a frame-by-frame processing manner instead of the video one, where $F = f$.

Upon training the autoencoder, a Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) is employed within the latent space to generate a denoised version of an input latent $\mathbf{z}_t$ at each timestep t. During denoising, the diffusion model can be conditioned on additional inputs, such as a text embedding $\mathbf{c} = f_{\text{CLIP}}\left(\mathbf{y}\right)$ generated by a pre-trained CLIP text encoder (Radford et al., 2021), corresponding to the input text prompt $\mathbf{y}$. The DDPM model $\epsilon_\theta(\cdot)$, a 3D U-Net parametrized by θ, optimizes the following loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}\sim\mathcal{E}(\mathbf{v}),\mathbf{c}=f_{\text{CLIP}}(\mathbf{y}),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{1}),t}\left[\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, \mathbf{c}, t\right)\|_2^2\right], \tag{1}$$

During inference, a latent variable $\mathbf{z}_T$ is sampled from the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$ and subjected to sequential denoising procedures of the DDPM to derive a refined latent $\mathbf{z}_0$. This denoised latent $\mathbf{z}_0$ is then fed into the decoder to synthesize the corresponding video $\mathcal{D}\left(\mathbf{z}_0\right)$.

Score distillation sampling. Score distillation sampling (SDS) (Poole et al., 2023; Wang et al., 2023a) employs the priors of pre-trained text-to-image models to facilitate text-conditioned 3D generation. Specifically, given a pre-trained diffusion model $\epsilon_\theta(\cdot)$ and the conditioning embedding $\mathbf{c} = f_{\text{CLIP}}\left(\mathbf{y}\right)$ corresponding to the text prompt $\mathbf{y}$, SDS optimizes a set of parameters ϕ of a differentiable parametric image generator $\mathcal{G}(\cdot)$ (*e.g.*, NeRF (Mildenhall et al., 2020)) using the gradient of the SDS loss $\mathcal{L}_{\text{SDS}}$:

$$\nabla_\phi \mathcal{L}_{\text{SDS}} = w(t)\left(\epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t) - \epsilon\right)\frac{\partial \mathbf{x}}{\partial \phi}, \tag{2}$$

where ϵ is sampled from $\mathcal{N}\left(\mathbf{0}, \mathbf{1}\right)$, $\mathbf{x}$ is an image rendered by $\mathcal{G}$, $\mathbf{z}_t$ is obtained by adding Gaussian noise ϵ to $\mathbf{x}$ corresponding to the timestep t of the diffusion process, $w(t)$ is a constant that depends on the noising schedule. Inspired by this method, we proposes a noise-invariant video score distillation sampling (NI-VSDS) strategy to efficiently harness the motion prior learned by the text-to-video diffusion model.

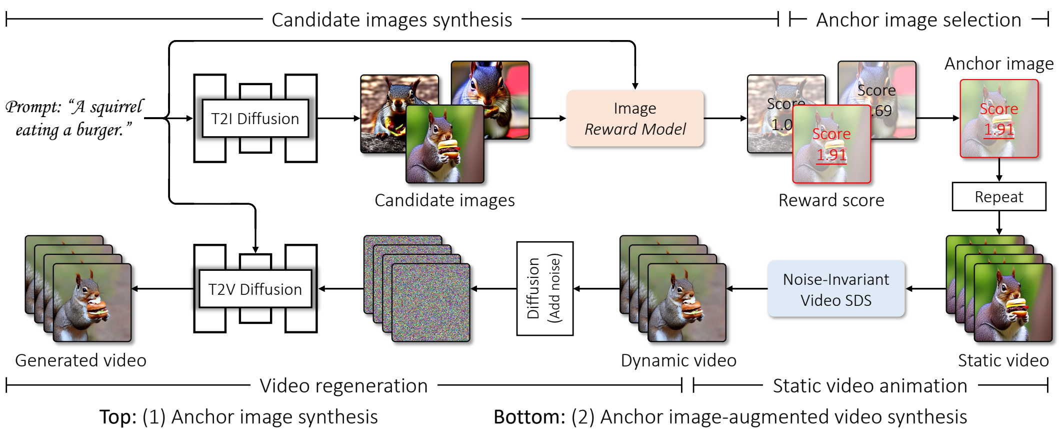


Figure 3: **Illustration of I4VGEN.** I4VGEN is a novel video diffusion inference pipeline, which enhances pre-trained text-to-video diffusion models by incorporating image reference information into the inference process. Instead of the vanilla text-to-video inference pipeline, I4VGEN consists of two stages: (1) anchor image synthesis and (2) anchor image-augmented text-to-video synthesis. Firstly, a simple yet effective generation-selection strategy is applied to synthesize candidate images and select the most suitable image using a reward-based mechanism, thereby obtaining high-quality anchor image. Subsequently, an innovative noise-invariant video scoring distillation sampling (NI-VSDS) is developed, which extracts motion prior from the text-to-video diffusion model to animate the anchor image into dynamic video, followed by a video regeneration process to refine the video.

3 I4VGEN

This section introduces **I4VGEN**, a novel video diffusion inference pipeline designed for enhancing the capabilities of pre-trained text-to-video diffusion models. As illustrated in Fig. 3, we factorize the inference process into two stages: (1) anchor image synthesis to generate the anchor image $\mathbf{x}$ given the text prompt $\mathbf{y}$, and (2) anchor image-augmented video synthesis to generate the video $\mathbf{v}$ by leveraging the text prompt $\mathbf{y}$ and the anchor image $\mathbf{x}$. This section provides the detailed explanations of both stages in Sec. 3.1 and 3.2, respectively.

3.1 ANCHOR IMAGE SYNTHESIS

The goal of this stage is to synthesize visually-realistic anchor images $\mathbf{x}$ that accurately correspond to the given text prompts $\mathbf{y}$. This image serves as a foundation to provide appearance information for enhancing the performance of the subsequent video generation. As illustrated in Fig. 3 (Top), a simple yet effective generation-selection pipeline is employed to produce the anchor image, which consist of candidate images synthesis and reward-based anchor image selection.

Candidate images synthesis. Instead of generating a single image, our approach produces a set of candidate images to ensure the selection of the best example. Utilizing a pre-trained image diffusion model $\mathcal{D}_{\text{img}}(\cdot)$, we construct the candidate image set as follows:

$$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N = \mathcal{D}_{\text{img}}\left(\mathbf{y}, \mathbf{z}_1\right), \mathcal{D}_{\text{img}}\left(\mathbf{y}, \mathbf{z}_2\right), \cdots, \mathcal{D}_{\text{img}}\left(\mathbf{y}, \mathbf{z}_N\right), \tag{3}$$

where N denotes the number of candidate images, and $\mathbf{z}_i$ represents Gaussian noise.

Reward-based anchor image selection. With the help of the image reward model $\mathcal{R}(\cdot)$ (Xu et al., 2023), a promising automatic text-to-image evaluation metric aligned with human preferences, the candidate image with the highest reward score s is selected as the anchor image $\mathbf{x}$, as defined by:

$$\mathbf{x} = \mathbf{x}_i, \quad \text{where } i = \arg\max_i s_i = \arg\max_i \mathcal{R}\left(\mathbf{x}_i\right). \tag{4}$$

The generation-selection design facilitates the acquisition of a high-quality anchor image, particularly beneficial for complex text prompts (see Fig. 5). Notably, our method accommodates both user-provided and retrieved images, extending its applicability to a variety of custom scenarios, as discussed in Sec. 4.5.

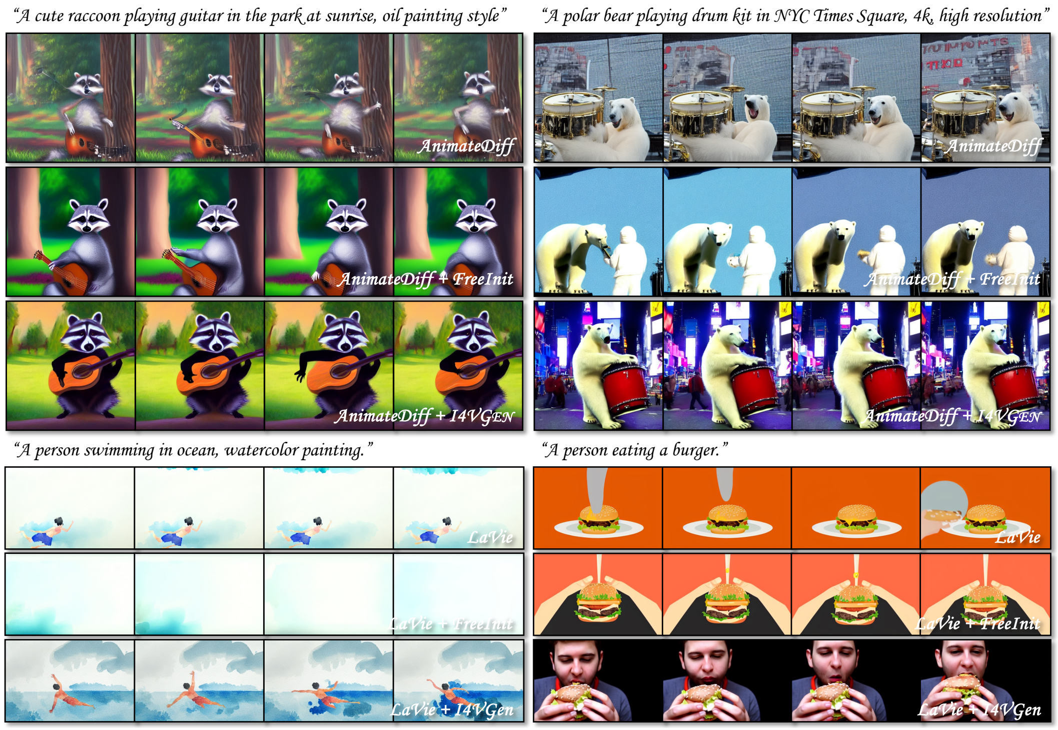


Figure 4: **Qualitative comparison.** Each video is generated with the same text prompt and random seed for all methods. Our approach significantly improves the quality of the generated videos while showing excellent alignment with text prompts.

3.2 Anchor image-augmented video synthesis

Upon obtaining the anchor image $\mathbf{x}$, we replicate it F times to create an initial static video $\hat{\mathbf{v}} \in \mathcal{V}, \hat{\mathbf{v}} = \{\mathbf{x}, \mathbf{x}, \cdots, \mathbf{x}\}$. The goal of this stage is to convert this static video into a high-quality video reflecting the text prompt $\mathbf{y}$. As illustrated in Fig. 3 (Bottom), we introduce static video animation and video regeneration.

Static video animation. A straightforward approach to animate the static video involves applying a diffusion-denoising process to transition from the static to dynamic state. However, this approach still encounters a training-inference gap, as the text-to-video diffusion model is trained on dynamic real-world videos but tested on static videos, leading to sub-optimal motion quality due to the introduction of static priors, as discussed in Sec. 4.4.

To address this limitation, we propose a novel approach leveraging the motion prior from the pre-trained text-to-video diffusion model to animate static videos. Drawing inspiration from score distillation sampling (SDS) as introduced in (Poole et al., 2023; Wang et al., 2023a), we develop the noise-invariant video score distillation sampling (NI-VSDS). Unlike vanilla SDS, which optimizes a parametric image generator, our approach directly parameterizes the static video $\hat{\mathbf{v}}$ and applies targeted optimization to it. The NI-VSDS loss function is defined as follows:

$$\nabla_{\hat{\mathbf{v}}} \mathcal{L}_{\text{NI-VSDS}} = w(t) \left(\epsilon_\theta(\hat{\mathbf{v}}_t, \mathbf{c}, t) - \epsilon \right), \tag{5}$$

where $\hat{\mathbf{v}}_t$ represents the noisy video at timestep t perturbed by Gaussian noise ϵ. Furthermore, we incorporate three strategic modifications:

- Instead of resampling the Gaussian noise at each iteration as in traditional SDS, we maintain a constant noise across the optimization, enhancing convergence speed.
- Optimization is confined to the initial stages of the denoising process, where noise levels are higher, focusing on dynamic information distillation.
- We implement a coarse-to-fine optimization strategy, evolving from high to low noise levels, specifically from timestep T to $\tau_{\text{NI-VSDS}}$, where $T > \tau_{\text{NI-VSDS}} > 0$. This approach stabilizes the optimization trajectory and yields superior motion quality.

Table 1: **VBench evaluation results per dimension.** This table compares the performance of I4VGEN with other counterparts across each of the 16 VBench dimensions.

Methods	Subj. Cons.	Back. Cons.	Tem. Flick.	Moti. Smo.	Dyna. Degr.	Aest. Qual.	Imag. Qual.	Obje. Class
AnimateDiff	87.11%	95.22%	95.99%	93.12%	**74.89%**	56.07%	64.29%	83.69%
+ FreeInit	90.45%	96.57%	96.89%	95.66%	70.17%	59.25%	63.51%	87.55%
+ I4VGEN	**95.17%**	**97.73%**	**98.51%**	**96.45%**	57.72%	**64.68%**	**66.18%**	**92.59%**
LaVie	91.65%	96.30%	98.03%	95.73%	**71.94%**	59.64%	65.13%	91.25%
+ FreeInit	92.32%	96.35%	98.06%	95.83%	71.11%	59.41%	63.89%	89.13%
+ I4VGEN	**94.12%**	**96.90%**	**98.55%**	**96.37%**	70.55%	**60.88%**	**66.55%**	**92.26%**

Methods	Mult. Obje.	Hum. Acti.	Color	Spat. Rela.	Scene	Appe. Style	Tem. Style	Over. Cons.
AnimateDiff	22.61%	90.40%	81.73%	31.55%	45.61%	24.40%	24.49%	25.71%
+ FreeInit	26.92%	93.00%	86.39%	30.71%	44.61%	23.98%	25.03%	25.61%
+ I4VGEN	**57.22%**	**95.80%**	**91.98%**	**45.20%**	**54.67%**	**25.07%**	**26.11%**	**28.01%**
LaVie	24.02%	94.80%	83.64%	26.27%	52.89%	23.67%	24.94%	27.25%
+ FreeInit	22.59%	94.20%	84.34%	27.46%	52.70%	23.61%	24.85%	26.89%
+ I4VGEN	**32.77%**	**96.20%**	**88.59%**	**33.81%**	**55.64%**	**24.35%**	**25.62%**	**27.68%**

The implementation of noise-invariant video score distillation sampling (NI-VSDS) algorithm is detailed in Algorithm 1, which outlines the process of converting a static video into a dynamic video using the defined NI-VSDS loss. Notably, we only perform a single update from timestep T to $\tau_{\text{NI-VSDS}}$, requiring fewer than 50 iterations, this is a significant reduction compared to the thousands of iterations typically required for text-to-3D synthesis in SDS. α is a scalar that defines the step size of the gradient update. We empirically set $\tau_{\text{NI-VSDS}} = \mathsf{Int}(T \times p_{\text{NI-VSDS}})$.

Algorithm 1: NI-VSDS

Input: T2V diffusion model $\epsilon_\theta(\cdot)$, text prompt $\mathbf{y}$, static video $\hat{\mathbf{v}}$, timestep $\tau_{\text{NI-VSDS}}$.
Output: Dynamic video.

Sampling $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$; $\mathbf{c} = f_{\text{CLIP}}(\mathbf{y})$
for $t = T, \cdots, \tau_{\text{NI-VSDS}}$ **do**
  $\hat{\mathbf{v}}_t \leftarrow \mathsf{AddNoise}(\hat{\mathbf{v}}, \epsilon, t)$
  $\nabla_{\hat{\mathbf{v}}} \mathcal{L}_{\text{NI-VSDS}} \leftarrow w(t)\,(\epsilon_\theta(\hat{\mathbf{v}}_t, \mathbf{c}, t) - \epsilon)$
  $\hat{\mathbf{v}} \leftarrow \hat{\mathbf{v}} - \alpha \cdot \nabla_{\hat{\mathbf{v}}} \mathcal{L}_{\text{NI-VSDS}}$
return $\hat{\mathbf{v}}$

Video regeneration. After animating the static video, we further enhance the appearance detail quality of the video through a diffusion-denoising process. This stage is not affected by the aforementioned training-inference gap, thereby achieving more refined generation results.

Notably, we can flexibly add noise up to any timestep τ_{re}, calculated as $\tau_{\text{re}} = \mathsf{Int}(T \times p_{\text{re}})$, followed by the corresponding denoising process. This strategy not only preserves the fine appearance textures but also reduces the required denoising steps, thus streamlining the video synthesis process and elevating the overall quality of the resulting video.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Implementation details. I4VGEN a novel video diffusion inference pipeline that leverages advanced image techniques to enhance pre-trained text-to-video diffusion models without requiring additional training, and can be seamlessly integrated into existing text-to-video diffusion models. To ascertain the efficacy and adaptability of I4VGEN, we apply it to two well-regarded text-to-video diffusion models: AnimateDiff (Guo et al., 2024b) and LaVie (Wang et al., 2023e).

For AnimateDiff, the `mm-sd-v15-v2` motion module[1], alongside Stable Diffusion v1.5, is utilized to synthesize 16 consecutive frames at a resolution of 512×512 pixels for evaluation. For LaVie, the

[1] `https://github.com/guoyww/AnimateDiff`

Table 2: **Ablation study.** Orange highlights generation-selection, while yellow highlights NI-VSDS.

Methods	Subj. Cons.	Back. Cons.	Tem. Flick.	Moti. Smo.	Dyna. Degr.	Aest. Qual.	Imag. Qual.	Obje. Class
AnimateDiff	87.11%	95.22%	95.99%	93.12%	**74.89%**	56.07%	64.29%	83.69%
+ I4VGEN (w/o gen.-sel.)	94.89%	97.80%	98.28%	96.99%	55.91%	62.23%	64.18%	90.95%
+ I4VGEN (w/o NI-VSDS)	**96.47%**	**98.82%**	**98.99%**	**97.56%**	28.24%	**65.17%**	65.52%	**92.66%**
+ I4VGEN	95.17%	97.73%	98.51%	96.45%	**57.72%**	64.68%	**66.18%**	92.59%

Methods	Mult. Obje.	Hum. Acti.	Color	Spat. Rela.	Scene	Appe. Style	Tem. Style	Over. Cons.
AnimateDiff	22.61%	90.40%	81.73%	31.55%	45.61%	24.40%	24.49%	25.71%
+ I4VGEN (w/o gen.-sel.)	40.68%	94.40%	90.55%	37.79%	53.72%	24.76%	26.03%	26.62%
+ I4VGEN (w/o NI-VSDS)	**62.84%**	94.80%	91.95%	**47.57%**	**55.80%**	24.88%	25.72%	27.91%
+ I4VGEN	57.22%	**95.80%**	**91.98%**	45.20%	54.67%	**25.07%**	**26.11%**	**28.01%**

`base-version`[2] is employed to generate 16 consecutive frames at 320×512 pixels for evaluation. All other inference details adhere to the original settings described in Guo et al. (2024b) and Wang et al. (2023e), respectively. Notably, both AnimateDiff and LaVie possess inherent text-to-image generation capabilities when excluding the motion module. To avoid introducing additional GPU storage requirements, we leverage their corresponding image versions for text-to-image generation in I4VGEN. For AnimteDiff, we empirically set $N = 16$, $p_{\text{NI-VSDS}} = 0.4$, $\alpha = 1$, and $p_{\text{re}} = 1$. For LaVie, we empirically set $N = 16$, $p_{\text{NI-VSDS}} = 0.4$, $\alpha = 1$, and $p_{\text{re}} = 0.8$. All experiments are conducted on a single NVIDIA V100 GPU (32 GB).

Benchmark. I4VGEN is assessed using VBench (Huang et al., 2024), a comprehensive benchmark that evaluates video generation models across 16 disentangled dimensions, which is more authoritative than FVD. These dimensions provide a detailed analysis of generation quality from two overarching perspectives: video quality[3], focusing on the perceptual quality of the generated videos, and video-condition consistency[4], assessing how well the generated videos align with the provided conditions.

4.2 QUALITATIVE COMPARISON

Fig. 4 presents a comparative analysis of our results against state-of-the-art counterparts using identical text prompts and random seeds. I4VGEN excels in enhancing both the temporal consistency and the frame-wise quality, alongside superior alignment with the text prompts. For instance, in the case of "playing guitar", AnimateDiff suffers from poor video quality, and FreeInit encounters an incomplete guitar in the middle of the video. In contrast, our method effectively addresses these issues, maintaining stable temporal consistency. Furthermore, while baseline methods struggle with accurate synthesis of all text-described components, *e.g.*, "NYC Times Square", I4VGEN generates videos that are visually realistic and closely aligned with the text prompts by utilizing anchor images obtained by the generation-selection strategy.

4.3 QUANTITATIVE COMPARISON

Objective evaluation. Following the protocols established by VBench, we evaluate I4VGEN in terms of both video quality and video-text consistency. As detailed in Table 1, I4VGEN outperforms all other approaches in temporal quality (higher background and subject consistency, less flickering, and better smoothness), frame-wise quality (higher aesthetic and imaging quality), and video-text

[2] `https://github.com/Vchitect/LaVie`

[3] Video quality includes 7 evaluation dimensions: Subject Consistency, Background Consistency, Temporal Flickering, Motion Smoothness, Dynamic Degree, Aesthetic Quality, and Imaging Quality. The first 5 evaluate temporal quality, and the last 2 evaluate frame-wise quality.

[4] Video-condition consistency includes 9 evaluation dimensions: Object Class, Multiple Objects, Human Action, Color, Spatial Relationship, Scene, Appearance Style, Temporal Style, Overall Consistency. The first 6 evaluate semantics, the 7 and 8-th evaluate style, and the 9-th evaluates overall consistency.

Figure 5: **Intermediate results visualization.** We provide visualizations of the candidate images with reward scores, the dynamic video, and the corresponding generated video.

consistency (greater semantics, style, and overall consistency). Although counterparts occasionally produce videos with more dynamic motion, they are often linked to inappropriate or excessive movements. I4VGEN strikes a more effective balance between motion intensity and overall video quality, which is further verified in the user study.

User study. We conduct a subjective user study involving 20 volunteers with expertise in image and video processing, with each participant answering 15 questions. Specifically, participants are asked to select the video with the highest quality across three dimensions: video quality, video-condition consistency, and overall score. As shown in Table 3, our approach outperforms the other methods favorably.

Table 3: **User study.**

Method	Video Quality	Vid.-Cond. Consistency	Overall score
AnimateDiff	6.00%	10.67%	6.33%
+ FreeInit	27.67%	15.67%	25.00%
+ I4VGEN	**66.33%**	**73.67%**	**68.67%**
LaVie	27.67%	21.33%	22.33%
+ FreeInit	22.67%	18.33%	19.67%
+ I4VGEN	**49.67%**	**60.33%**	**58.00%**

Inference time. We define the time cost of a single denoising iteration for a video in a video diffusion model as c. For AnimateDiff (Guo et al., 2024b), following the original inference setting, the time cost to generate a single 16-frame video is $25c$. FreeInit requires 5 rounds of diffusion-denoising to generate a single video, taking a time of $5 \times 25c = 125c$. The time cost for I4VGEN to generate a single video is: $< 25c$ (for synthesizing 16 candidate images) $+ 0.6 \times 25c$ (for NI-VSDS) $+ \leq 25c$ (for video regeneration) $= < 65c$ (**total cost**), making it more efficient compared to FreeInit. LaVie (Wang et al., 2023e) shares the same conclusion.

Table 4: **Inference time.**

Method	Time
AnimateDiff	21.73s
AnimateDiff + FreeInit	113.67s
AnimateDiff + I4VGEN	53.78s

We also provide the inference time for a single video in Table 4, evaluated on a single NVIDIA V100 GPU (32 GB), where 50 videos are randomly generated to obtain an average inference time. Our method performs better than FreeInit.

4.4 ABLATION STUDY

On generation-selection strategy. We adopt a generation-selection strategy to create visually-realistic and semantically-faithful anchor images, which serve as a foundation for providing appearance information to enhance subsequent video generation performance. As shown in Table 2, highlighted in orange, compared to randomly synthesizing a single anchor image, the generation-selection strategy significantly improves the quality of the generated videos in terms of frame-wise quality and consistency with the text. Fig. 5 provides a visualization of the candidate images, where the reward-based selection strategy eliminates unsatisfactory images, leading to better results.

On NI-VSDS. Directly applying the video regeneration process to static videos introduces static priors, resulting in suboptimal motion quality. As shown in Table 2, highlighted in yellow, while direct diffusion-denoising improves the temporal consistency of the generated videos, it severely sacrifices the motion dynamics, adversely affecting the motion style. In contrast, our method achieves an effective balance between motion intensity and overall video quality.

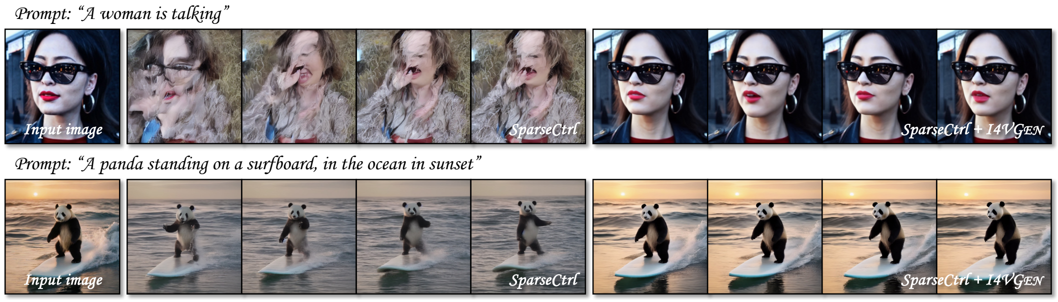


Figure 6: **Adaptation on SparseCtrl.** I4VGEN can be seamlessly integrated into SparseCtrl by replacing the anchor image with the provided image, leading to improved results.

On video regeneration. Fig. 5 visualizes the intermediate results, demonstrating that the video regeneration process is essential for refining appearance details.

4.5 MORE APPLICATIONS

Adaptation on real image. Our method adapts to user-provided images, as shown in Fig. 7, where we use real images as anchor images, resulting in high-fidelity videos that are semantically consistent with the real images. Notably, our approach differs from vanilla image-to-video generation, as the synthesized videos are not completely aligned with the provided images. NI-VSDS is designed to animate static videos and is implemented as a spatio-temporal co-optimization.

Figure 7: **Adaptation on real image.**

Adaptation on image-to-video diffusion models. I4VGEN can be seamlessly integrated into existing image-to-video diffusion models by replacing the anchor images with the provided images, thereby enhancing the overall video quality. As shown in Fig. 6, integrating I4VGEN into SparseCtrl (Guo et al., 2023) significantly improves the quality of the generated videos in terms of temporal consistency and appearance fidelity.

5 CONCLUSION

The paper introduces I4VGEN, a novel video diffusion inference pipeline to leverage advanced image techniques to enhance pre-trained text-to-video diffusion models, which requires no additional learnable parameters and training costs. I4VGEN decomposes the text-to-video inference process into anchor image synthesis and anchor image-augmented video synthesis. Correspondingly, a simple yet effective generation-selection strategy is applied to produce a high-quality anchor image, and an innovative noise-invariant video score distillation sampling (NI-VSDS) is designed to animate the image, followed by a video regeneration process to enhance the final output. I4VGEN effectively alleviates non-zero terminal signal-to-noise ratio issues and demonstrates improved visual realism and textual fidelity when integrated with existing video diffusion models.

Limitation and discussion. I4VGEN improves the video diffusion model but requires more inference cost. As discussed in Sec. 4.3, the inference time of I4VGEN is over double the baseline. Enhancing inference efficiency remains a future goal, with distillation techniques as a potential approach. Furthermore, removing the generation-selection strategy can reduce inference costs to some extent. As shown in Table 2, our method still significantly outperforms the baseline under this setting. Additionally, although our method and FreeInit are orthogonal, integrating both by replacing video regeneration with FreeInit fails to produce notable benefits.

REFERENCES

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

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

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023b.

Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A Efros, and Tero Karras. Generating long videos of dynamic scenes. In *NeurIPS*, 2022.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023a.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wave-grad: Estimating gradients for waveform generation. In *ICLR*, 2021.

Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023b.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *ICLR*, 2024a.

Xuweiyi Chen, Tian Xia, and Sihan Xu. Unictrl: Improving the spatiotemporal consistency of text-to-video diffusion models via training-free unified attention control. *arXiv preprint arXiv:2403.02332*, 2024b.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, 2023.

Martin Nicolas Everaert, Athanasios Fitsios, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Exploiting the signal-leak bias in diffusion models. In *WACV*, 2024.

Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, et al. Dreamoving: A human video generation framework based on diffusion models. *arXiv preprint arXiv:2312.05107*, 2023.

Gereon Fox, Ayush Tewari, Mohamed Elgharib, and Christian Theobalt. Stylevideogan: A temporal generative model using a pretrained stylegan. In *BMVC*, 2021.

Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In *CVPR*, 2023.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, 2022.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *ICCV*, 2023.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *ICLR*, 2024.

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Josh Susskind, and Navdeep Jaitly. Matryoshka diffusion models. *arXiv preprint arXiv:2310.15111*, 2023.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *CVPR*, 2024a.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024b.

Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.

Nicholas Guttenberg. Diffusion with offset noise. `https://www.crosslabs.org/blog/diffusion-with-offset-noise`.

Jiawei He, Andreas Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. Probabilistic video generation using holistic attribute control. In *ECCV*, 2018.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022b.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2021.

Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiangyang Li, Tao Qin, et al. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. In *NeurIPS*, 2022.

Mingxiao Li, Tingyu Qu, Wei Sun, and Marie-Francine Moens. Alleviating exposure bias in diffusion models through sampling with shifted time steps. In *ICLR*, 2024.

Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023.

Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *AAAI*, 2018.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *WACV*, 2024.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *AAAI*, 2022.

Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023.

Shijie Ma, Huayi Xu, Mengjian Li, Weidong Geng, Meng Wang, and Yaxiong Wang. Optimal noise pursuit for augmenting text-to-video generation. *arXiv preprint arXiv:2311.00949*, 2023.

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

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *ACM MM*, 2017.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *ICML*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017.

Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *IJCV*, 2020.

Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Mostgan-v: Video generation with temporal motion styles. In *CVPR*, 2023.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *ICLR*, 2021.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *ICLR*, 2023.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NIPS*, 2016.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023a.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023b.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023c.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023d.

Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *CVPR*, 2020.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023e.

Yuhan Wang, Liming Jiang, and Chen Change Loy. Styleinv: A temporal style modulated inversion network for unconditional video generation. In *ICCV*, 2023f.

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 2022.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan 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. In *ICCV*, 2023a.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. *arXiv preprint arXiv:2312.07537*, 2023b.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *ECCV*, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023.

Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. In *NeurIPS*, 2023.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Jaehoon Yoo, Semin Kim, Doyup Lee, Chiheon Kim, and Seunghoon Hong. Towards end-to-end generative modeling of long videos with memory-efficient bidirectional transformers. In *CVPR*, 2023.

Jiwen Yu, Xiaodong Cun, Chenyang Qi, Yong Zhang, Xintao Wang, Ying Shan, and Jian Zhang. Animatezero: Video diffusion models are zero-shot image animators. *arXiv preprint arXiv:2312.03793*, 2023a.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, 2023b.

Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *ICLR*, 2022.

Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *CVPR*, 2023c.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023a.

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

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

This appendix is structured as follows:

- In Appendix A, we provide a discussion of related work.
- In Appendix B, we provide additional experiment results and analysis.
- In Appendix C, we provide the code for I4VGEN.

A RELATED WORK

Video Generative Models. The domain of video generation has seen significant advancements through the use of Generative Adversarial Networks (GANs) (Vondrick et al., 2016; Saito et al., 2017; Tulyakov et al., 2018; Wang et al., 2020; Saito et al., 2020; Tian et al., 2021; Fox et al., 2021; Yu et al., 2022; Skorokhodov et al., 2022; Brooks et al., 2022; Shen et al., 2023; Wang et al., 2023f), Variational Autoencoders (VAEs) (Mittal et al., 2017; Li et al., 2018; He et al., 2018), and Autoregressive models (ARs) (Yan et al., 2021; Ge et al., 2022; Wu et al., 2022; Hong et al., 2023; Villegas et al., 2023; Fu et al., 2023; Yoo et al., 2023; Yu et al., 2023b). Despite these developments, synthesizing videos from text prompts remains challenging due to the complexities of modeling spatio-temporal dynamics. Recent innovations driven by the successes of diffusion models (Ho et al., 2020; Dhariwal & Nichol, 2021; Song et al., 2021), which have been applied effectively in image generation (Rombach et al., 2022; Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Gu et al., 2023; Balaji et al., 2022; Xue et al., 2023; Meng et al., 2022; Guo et al., 2024a) and audio synthesis (Kong et al., 2021; Chen et al., 2021; Popov et al., 2021; Leng et al., 2022; Liu et al., 2022), and underscore the emergence of substantial headway (Ho et al., 2022b;a; He et al., 2022; Singer et al., 2023; Blattmann et al., 2023b; Yu et al., 2023c; Ruan et al., 2023; Wu et al., 2023a; Chen et al., 2023a;b; Esser et al., 2023; Ge et al., 2023; Chen et al., 2024a; Geyer et al., 2024; Ma et al., 2023; Wang et al., 2023e; Zhang et al., 2023a;b; Hu et al., 2023; Wang et al., 2023d; Feng et al., 2023; Guo et al., 2024b; Girdhar et al., 2023; Blattmann et al., 2023a; Gupta et al., 2023; Wang et al., 2023b;c; Luo et al., 2023) in research endeavors devoted to video synthesis from text input.

The foundational contributions of the Video Diffusion Model (VDM) (Ho et al., 2022b) represents a milestone in leveraging diffusion models for video generation by adapting the 2D U-Net architecture used in image generation to a 3D U-Net capable of temporal modeling. Successive researches, such as Make-A-Video (Singer et al., 2023) and Imagen Video (Ho et al., 2022a), expand video generation capabilities significantly. To enhance efficiency, subsequent models have transitioned the diffusion process from pixel to latent space (He et al., 2022; Zhou et al., 2022; Wang et al., 2023b; Blattmann et al., 2023b;a; Guo et al., 2024b; Wang et al., 2023e), paralleling advancements in latent diffusion for images (Rombach et al., 2022).

However, the direct generation of videos from text prompts remains intrinsically challenging. Recent approaches (Blattmann et al., 2023a; Zhang et al., 2023b; Girdhar et al., 2023; Chen et al., 2023a; 2024a; Li et al., 2023; Hu et al., 2023; Yu et al., 2023a; Ren et al., 2024) have employed text-to-image synthesis as an intermediary step, enhancing overall performance. Despite these advancements, these methods still face the challenge of high computational training costs. In this study, we explore a novel training-free methodology aimed at bridging the existing gap in the field.

In addition, (Chen et al., 2024b) (contemporary researches) introduces additional operations in the attention layer, *i.e.*, cross-frame self-attention control, to enhance the video model. However, this necessitates modifications to the model architecture, whereas our method does not.

Signal-Leak Bias. Diffusion models are designed to generate high-quality visuals from noise through a sequential denoising process, which is consistent in both image and video diffusion models. During training, Gaussian noise corrupts the visual content, challenging the model to restore it to its original form. In the inference phase, the model operates on pure Gaussian noise, transforming it into a realistic visual content step-by-step.

Unfortunately, most existing diffusion models exhibit a disparity between the corrupted image during training and the pure Gaussian noise during inference. Commencing denoising from pure Gaussian noise in the inference phase deviates from the training process, potentially introducing *signal-leak bias*. For image diffusion models, (Guttenberg; Lin et al., 2024; Li et al., 2024) point out flaws in common diffusion noise schedules and sample steps, and propose to fine-tune the diffusion model to mitigate or eliminate the signal-leak bias during training, leading to improved results. (Everaert et al., 2024) attempts to exploit signal-leak bias to achieve more control over the generated images. For video diffusion models, this issue becomes more pronounced. (Wu et al., 2023b; Ma et al., 2023) invert the retrieved video or generated low-quality to construct initial noise to alleviate the problem of signal-leak, improving inference quality. However, they suffer from limited diversity and cumbersome inference. At the same time, first-round inference of FreeInit (Wu et al., 2023b) still exhibits a training-inference gap.

In contrast to existing methods, our approach utilizes images as the stepping stone for text-to-video generation. This novel pathway aims to produce visually-realistic and semantically-reasonable videos while maintaining manageable computational overheads, as detailed in Sec. 4.

B EXPERIMENTS

B.1 QUALITATIVE COMPARISON

We provide more visualization results in Fig. 8, it can be seen that our method generates more semantically plausible and photo-realistic results than its counterparts. We provide the videos shown in the main paper and appendix in `mp4` format in the Supplementary material.

B.2 QUANTITATIVE COMPARISON

On hyperparameters. I4VGEN is a training-free method that improves video generation performance by correcting the inference process. It is obvious that I4VGEN is also a case-wise method, where different cases correspond to different optimal hyperparameters. In this paper, we provide an empirical setting that is mild for most instances, serving as a performance lower bound for I4VGEN, and facilitating large-scale quantitative comparisons. Furthermore, we also provide a visualization of the impact of hyperparameters in Fig. 9, which shows that carefully tuned hyperparameters can achieve higher-quality videos.

B.3 FAILURE CASES AND DISCUSSIONS

We provide the failure cases in Fig. 10, I4VGEN is designed to fully unleash the potential of existing video diffusion models, but it still fails to synthesize high-quality videos that are out of the distribution.

C CODE

We also provide the code for I4VGEN in the Supplementary material.

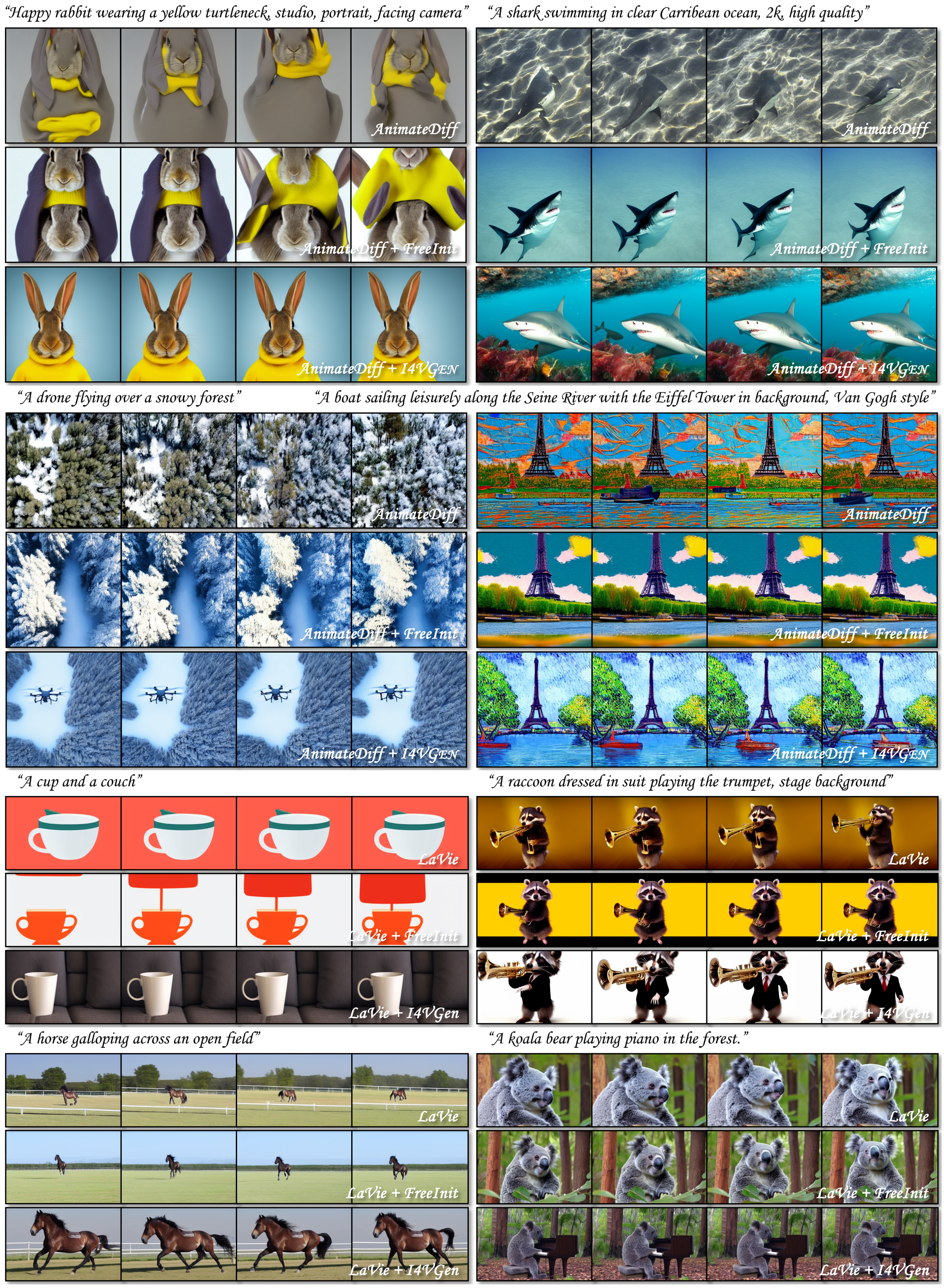


Figure 8: **Qualitative comparison.** Each video is generated with the same text prompt and random seed for all methods. Our approach significantly improves the quality of the generated videos while showing excellent alignment with text prompts.

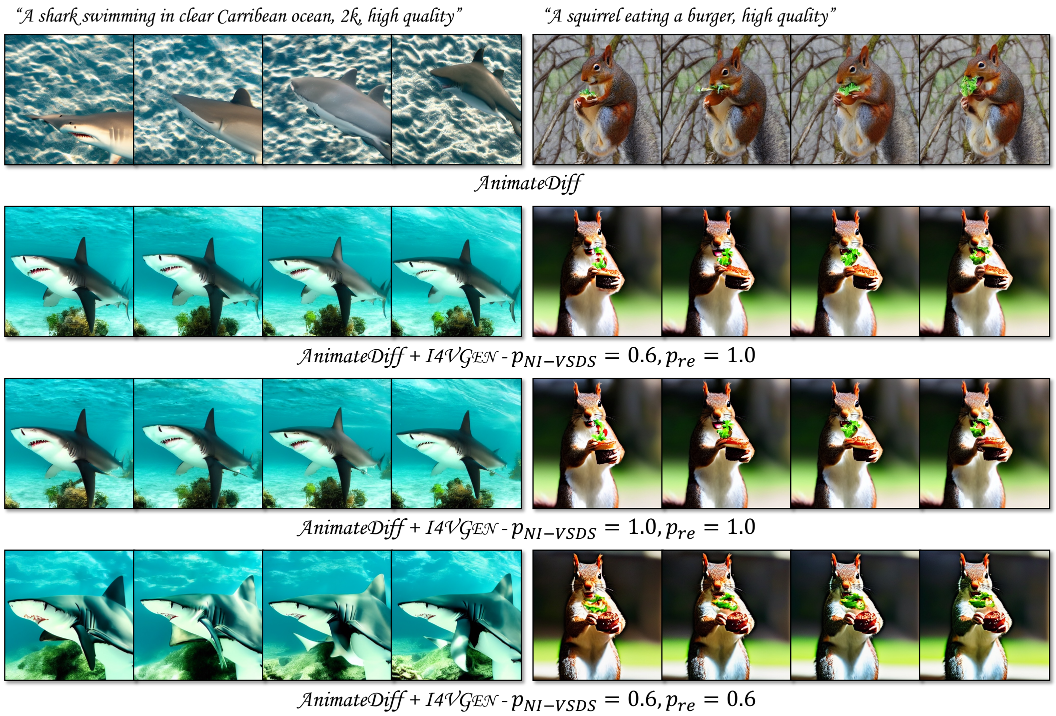


Figure 9: **Impact of hyperparameters.** For different texts, the optimal parameter settings are different, and the sensitivity to parameters also varies. However, they all significantly outperform the baseline. In this paper, we provide an empirical setting that is mild for most cases, serving as a performance lower bound for I4VGEN. I4VGEN supports fine-tuning parameters on a per-example, achieving higher-quality videos.

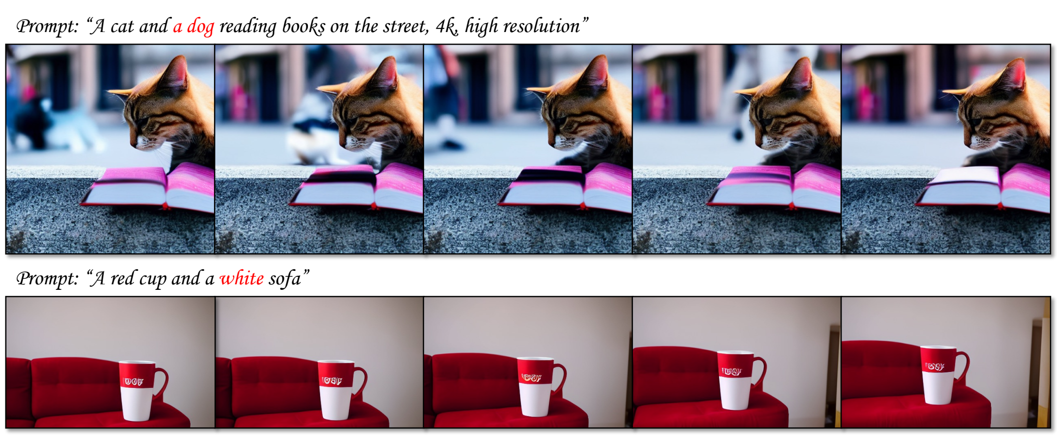


Figure 10: **Failure cases.** I4VGEN is designed to fully unleash the potential of existing video diffusion models, but it still cannot synthesize high-quality videos that are out of the distribution. For example, the text marked in red.

D ADDITIONAL EXPERIMENTAL RESULTS

D.1 ADDITIONAL VISUAL RESULTS ON DYNAMICRAFTER

We integrate I4VGEN into DynamiCrafter (Xing et al., 2024), which exhibits state-of-the-art performance on the VBench Image-to-Video Leaderboard. As shown in Fig. 11, the beginnings and endings of videos generated by DynamiCrafter suffer from low quality. For example, the front part of the face video generated by DynamiCrafter exhibits serious artifacts, and the face in the latter part are deformed. Our method alleviates these issues, which demonstrates that I4VGEN can significantly improve the quality of videos synthesized by DynamiCrafter.

D.2 ADDITIONAL VISUAL RESULTS ON SPARSECTRL

We conduct experiments on action instructions. As shown in Fig. 12, we explore two prompt-based motion enhancement strategies:

- By providing static descriptions in negative prompt, the dynamic intensity of the synthesized videos can be further enhanced.
- By providing specific action instruction in the prompt, such as "waving its hands", the synthesized video accurately renders this action.

These findings indicate that I4VGEN does not compromise the dynamic nature of the synthesized videos but rather depicts more reasonable and accurate motion.

D.3 ADDITIONAL VISUAL RESULTS USING FLUX

We provide the visual results of I4VGEN adapted on FLUX in the Fig. 13. Despite the detailed and realistic images synthesized by FLUX, AnimateDiff + I4VGEN is still constrained by the video baseline, *i.e.*, AnimateDiff, in rendering image details and is unable to synthesize realistic videos. Evidently, the distribution of images synthesized by FLUX exceeds what AnimateDiff can handle, which relies on SD 1.5. However, the layout and composition information of images synthesized by FLUX still provide strong support for video synthesis, resulting in promising outcomes.

D.4 ADDITIONAL VISUAL RESULTS ON ANIMATEDIFF

We provide the visual results on AnimateDiff using the `Realistic Vision V5.1` LoRA in the Fig. 14. Our method still significantly improves the quality of the generated videos while showing excellent temporal consistency.

D.5 ADDITIONAL INTERMEDIATE RESULTS VISUALIZATION

We provide additional intermediate results in the Fig. 17.

D.6 ADDITIONAL VISUAL RESULTS ON LARGE CAMERA POSE CHANGE

We provide more visual results involving significant changes in camera poses in Fig. 16, which demonstrate that our method can handle this scenario, improving the temporal consistency and smoothness of the synthesized videos.

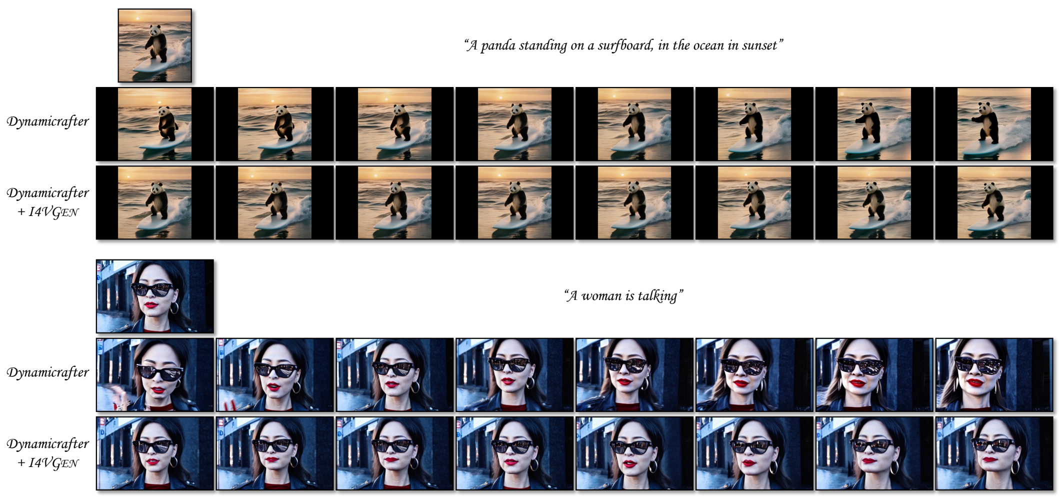


Figure 11: **Additional visual results on DynamiCrafter.** We provide the videos in `mp4` format in the supplementary material for better viewing.

Figure 12: **Additional visual results on SparseCtrl.** We provide the videos in `mp4` format in the supplementary material for better viewing.

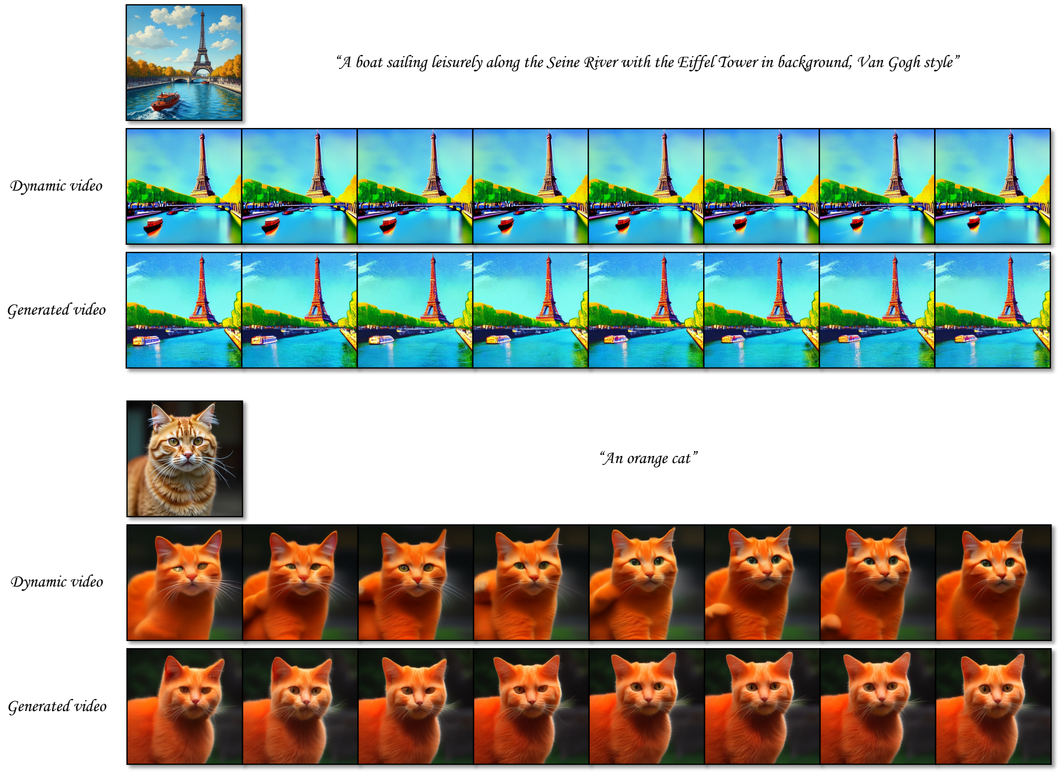


Figure 13: **Adaptation on image generated by FLUX.** We provide the videos in `mp4` format in the supplementary material for better viewing.

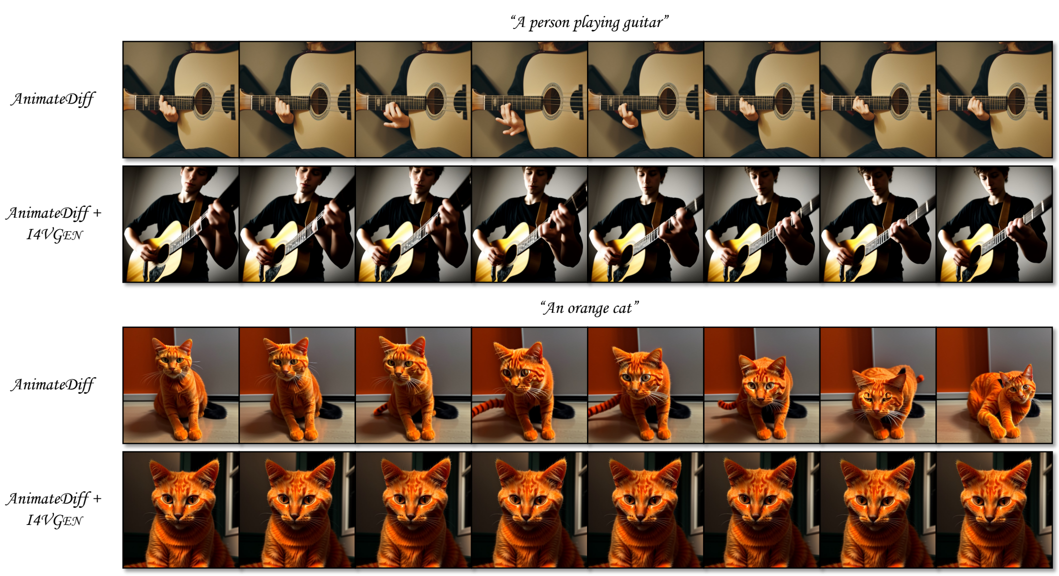


Figure 14: **Additional visual results on AnimateDiff using LoRA.** We provide the videos in `mp4` format in the supplementary material for better viewing.

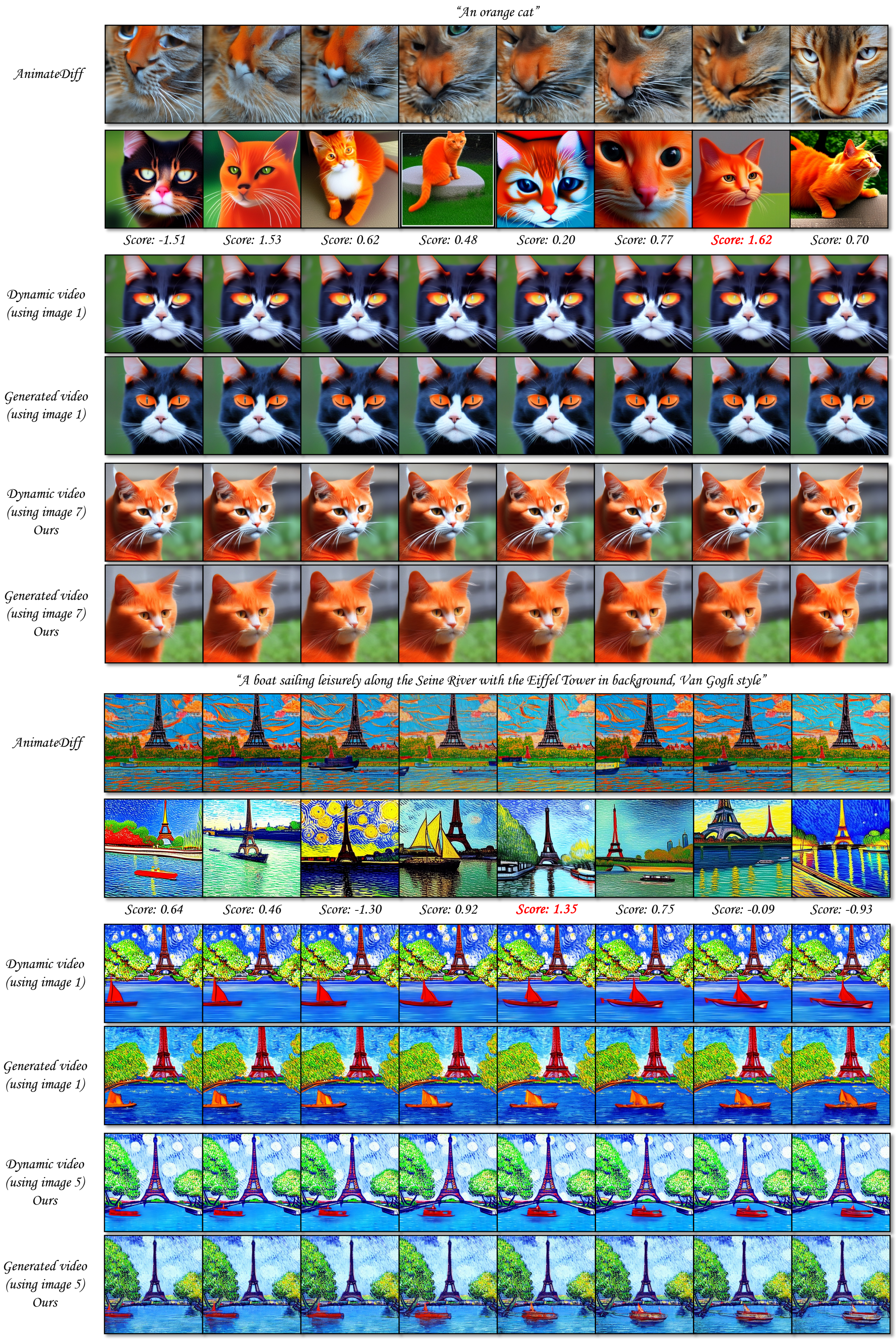


Figure 15: **Additional intermediate results visualization.** We provide the videos in `mp4` format in the supplementary material for better viewing.

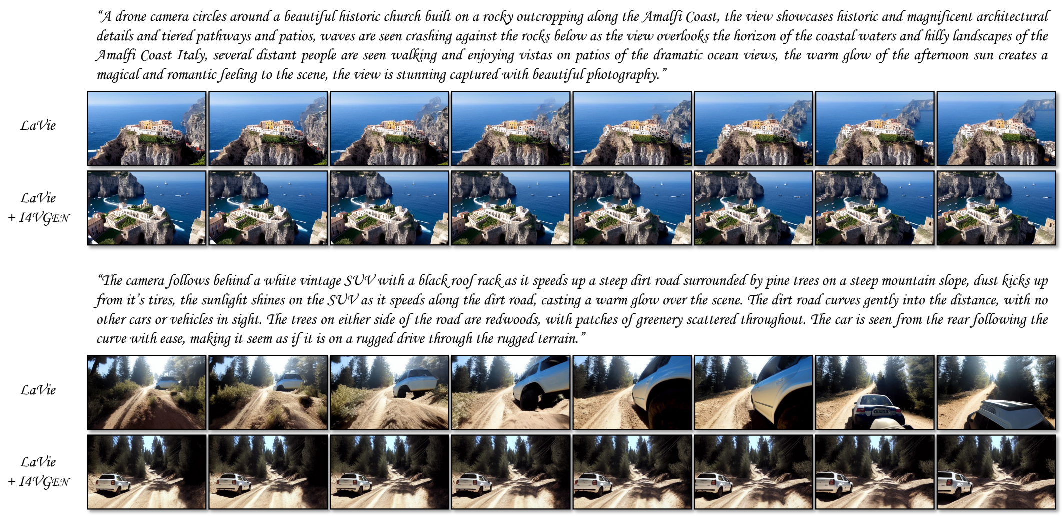


Figure 16: **Additional visual results on large camera pose change.** We provide the videos in `mp4` format in the supplementary material for better viewing.

Figure 17: **Impact of p_{re}.** We provide the videos in `mp4` format in the supplementary material for better viewing.