VideoElevator: Elevating Video Generation Quality with Versatile Text-to-Image Diffusion Models

Yabo Zhang[1], Yuxiang Wei[1], Xianhui Lin[2], Zheng Hui[2], Peiran Ren[2], Xuansong Xie[2], Wangmeng Zuo[1]*

[1] Harbin Institute of Technology
[2] Tongyi Lab
{hitzhangyabo2017,yuxiang.wei.cs,xhlin129,renpeiran,xiexuansong}@gmail.com,
zheng_hui@aliyun.com, wmzuo@hit.edu.cn

Abstract

Text-to-image diffusion models (T2I) have demonstrated unprecedented capabilities in creating realistic and aesthetic images. On the contrary, text-to-video diffusion models (T2V) still lag far behind in frame quality and text alignment, owing to insufficient quality and quantity of training videos. In this paper, we introduce VideoElevator, a *training-free* and *plug-and-play* method, which elevates the performance of T2V using superior capabilities of T2I. Different from conventional T2V sampling (*i.e*, temporal and spatial modeling), VideoElevator explicitly decomposes each sampling step into *temporal motion refining* and *spatial quality elevating*. Specifically, temporal motion refining uses encapsulated T2V to enhance temporal consistency, followed by inverting to the noise distribution required by T2I. Then, spatial quality elevating harnesses inflated T2I to directly predict less noisy latent, adding more photo-realistic details. We have conducted experiments in extensive prompts under the combination of various T2V and T2I. The results show that VideoElevator not only improves the performance of T2V baselines with foundational T2I, but also facilitates stylistic video synthesis with personalized T2I.

Code — https://github.com/YBYBZhang/VideoElevator
Project — https://videoelevator.github.io/

1 Introduction

Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song et al. 2021) have facilitated the rapid development of generative modeling, and demonstrated tremendous success in multiple modalities (Poole et al. 2022; Ramesh et al. 2022; Ho et al. 2022a,b; Singer et al. 2023; Blattmann et al. 2023b), especially in image and video synthesis. Nonetheless, exceptional generative capabilities primarily depend on an abundance of high-quality datasets. For example, text-to-image diffusion models (T2I) requires billions of highly aesthetic images to achieve desirable control and quality (Ramesh et al. 2022; Saharia et al. 2022; Rombach et al. 2022; Podell et al. 2023). However, owing to the difficulty in collection, the publicly available video datasets (Bain et al. 2021) are far behind in both quantity and quality (*i.e*, million-level scale and low-quality contents), which greatly limits the text-to-video diffusion models (T2V) (Guo et al. 2023; Wang et al. 2023a,c; Zhang et al. 2023a) in prompt fidelity and frame quality.

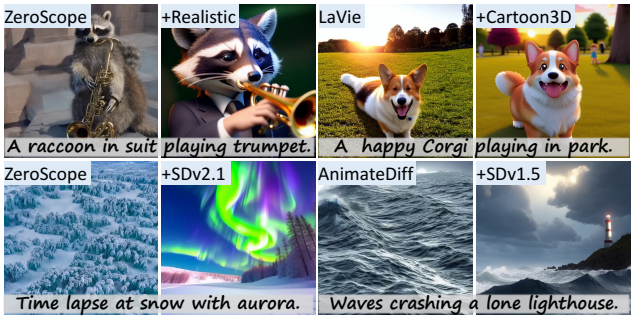

Figure 1: Videos enhanced by VideoElevator. VideoElevator aims at elevating the quality of videos generated by existing text-to-video models (*e.g*, ZeroScope) with text-to-image diffusion models (*e.g*, RealisticVision). It supports cooperation of various image and video diffusion models.

Recent studies (Ho et al. 2022b,a; Wang et al. 2023a; Guo et al. 2023; Wang et al. 2023c; Zhang et al. 2023a) promisingly enhance video generation performance through the advances in T2I. They either jointly train a T2V using video and image datasets (Ho et al. 2022b,a), or initialize it with a pre-trained T2I (Wang et al. 2023a; Guo et al. 2023; Wang et al. 2023c; Zhang et al. 2023a). With the spatial modeling capabilities of T2I, T2V can concentrate more on the learning of temporal dynamics. However, as training progresses, T2V inevitably will be biased towards low visual quality caused by the training videos (Bain et al. 2021), such as lower resolutions, blurriness, and inconsistent descriptions. Even when integrating with higher-quality personalized T2I, AnimateDiff (Guo et al. 2023) cannot rectify the bias towards low visual quality in temporal layers. As a result, synthesized videos still suffer from the frame quality degradation issue in comparison to T2I generated images, and sometimes contain visible flickers.

Considering that low-quality contents accumulate throughout T2V sampling chain, we investigate to rectify

*Corresponding Author

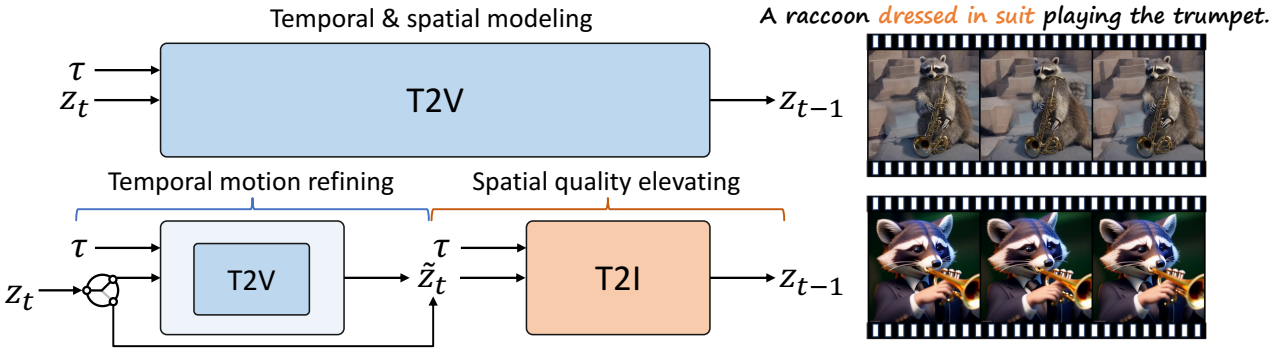

Figure 2: VideoElevator for improved text-to-video generation. VideoElevator explicitly decomposes each step into temporal motion refining and spatial quality elevating, where the former encapsulates T2V to enhance temporal consistency and the latter harnesses T2I to provide more faithful details, *e.g*, dressed in suit.

it with high-quality T2I at each step. As shown in Fig. 2, we introduce *VideoElevator* to explicitly decompose each sampling step into two components: (i) temporal motion refining *and* (ii) spatial quality elevating. Temporal motion refining encapsulates T2V to produce more temporally-consistent video latent and then inverts the latent to current timestep, so that spatial quality elevating can directly leverage T2I to add realistic details. Compared to vanilla T2V sampling, decomposed sampling potentially raises the bar of synthesized video quality.

We implement *VideoElevator* in a *plug-and-play* manner to elevate the performance of text-to-video diffusion models with various text-to-image diffusion models. Specifically, given video latent z_t at timestep t, *temporal motion refining* encapsulates T2V with a temporal low-pass frequency filter to improve consistency, followed by T2V-based SDEdit (Meng et al. 2021) to portray natural motion. To obtain noise latent $\tilde{z}_t$ required by T2I, it deterministically inverts (Song, Meng, and Ermon 2021) video latent to preserve motion as much as possible. After that, *spatial quality elevating* leverages inflated T2I to directly convert $\tilde{z}_t$ to less noisy z_{t-1}, where the self-attention in T2I is extended into cross-frame attention for appearance consistency. Empirically, applying temporal motion refining in several timesteps is sufficient to ensure temporal consistency, so we omit it in other timesteps for efficiency. To ensure interaction between various T2V and T2I, VideoElevator projects all noise latents to clean latents before being fed into another model. Benefiting from it, VideoElevator supports the combination of various T2V and T2I, as long as their clean latent distributions are shared (*i.e*, same autoencoders).

Evaluating on extensive video prompts, VideoElevator not only greatly improves the generation quality of T2V with foundational T2I, but also facilitates creative video synthesis with versatile personalized T2I. Firstly, it visibly enhances a wide range of T2V with Stable Diffusion V1.5 or V2.1-base (Rombach et al. 2022), achieving higher frame quality and prompt consistency than their baselines. Secondly, it enables T2V to replicate the diverse styles of personalized T2I more faithfully than AnimateDiff (Guo et al. 2023), while supporting T2I with significant parameter shifts, *e.g*, DPO-enhanced T2I (Wallace et al. 2023).

To summarize, our key contributions are as three-fold:

- We introduce VideoElevator, a *training-free* and *plug-and-play* method, which enhances quality of synthesized videos with versatile text-to-image diffusion models.
- To enable cooperation between various T2V and T2I models, we decompose each sampling step into *temporal motion refining* and *spatial quality elevating*, where the former applies encapsulated T2V to improve temporal consistency and the latter harnesses inflated T2I to provide high-quality details.
- The experiments show that VideoElevator significantly improves the performance of T2V baselines by leveraging versatile T2I, in terms of frame quality, prompt consistency, and aesthetic styles.

2 Related Work

2.1 Text-to-image diffusion models

Foundational text-to-image diffusion models. Diffusion models have achieved unprecedented breakthroughs in image creation and editing tasks (Ramesh et al. 2022; Saharia et al. 2022; Rombach et al. 2022; Podell et al. 2023; Zhang, Rao, and Agrawala 2023), significantly surpassing the performance of previous models based on GANs (Sauer et al. 2023) and auto-regressive models (Ramesh et al. 2021; Yu et al. 2022). Benefiting from large-scale exceptional training data, foundational text-to-image diffusion models can producing images with photo-realistic quality and textual mastery. To reduce computational complexity, Latent Diffusion Model (LDM) (Rombach et al. 2022) applies diffusion process in the latent space of auto-encoders. Stable Diffusion (SD) is the most popular open-sourced instantiation of LDM, and facilitates a surge of recent advances in downstream tasks (Zhang, Rao, and Agrawala 2023; Gal et al. 2023; Poole et al. 2022; Ma et al. 2023a; Huang et al. 2023a; Lin et al. 2023; Ni et al. 2023; Lv et al. 2024). By increasing parameters and improving training strategies of SD, Stable Diffusion XL (SDXL) (Podell et al. 2023) obtains drastically better performance than the original Stable Diffusion.

Personalized text-to-image diffusion models. Adapting from foundational text-to-image models, personalized text-to-image models (Gal et al. 2023; Ruiz et al. 2022; Wei et al. 2023; Cai et al. 2023; Wallace et al. 2023) can satisfy a wide range of user requirements, *e.g*, various aesthetic styles and human preference alignment. In particular, DreamBooth (Ruiz et al. 2022) and LoRA (Hu et al. 2021) enable users to efficiently finetune foundation models on their customized small datasets, thereby promoting the release of various stylistically personalized models. Diffusion-DPO (Wallace et al. 2023) refines on large-scale preference benchmarks and aligns Stable Diffusion better with human preference. Among them, DreamBooth and LoRA only slightly changes the parameters of foundation models, while Diffusion-DPO greatly alter their parameters.

2.2 Text-to-video diffusion models

Direct text-to-video diffusion models. Most text-to-video diffusion models (Chen et al. 2023; Wang et al. 2023b; He et al. 2022; Wang et al. 2023a,c; Zhang et al. 2023a; Esser et al. 2023; Ma et al. 2023b; Khachatryan et al. 2023; Zhang et al. 2023c; Wu et al. 2023a; Ge et al. 2023; Liang et al. 2023; Bar-Tal et al. 2024) are derived from text-to-image diffusion models, often incorporating additional temporal layers. VDM (Ho et al. 2022b) and Image-Video (Ho et al. 2022a) jointly train models from the scratch using large-scale video-text and image-text pairs. Based on pre-trained Stable Diffusion, LaVie (Wang et al. 2023c) and ZeroScope (Wang et al. 2023a) finetune the whole model during training. Differently, VideoLDM (Blattmann et al. 2023b) and AnimateDiff (Guo et al. 2023) only finetune the additional temporal layers, enabling them to be plug-and-play with personalized image models. We note that Stable Diffusion based video models frozen the pre-trained autoencoders and thus share the same clean latent distribution. Compared to AnimateDiff (Guo et al. 2023), our VideoElevator produces higher-quality videos and supports a wider range of personalized image models.

Factorized text-to-video diffusion models. Due to the low-quality contents in training videos (Chen et al. 2023; Wang et al. 2023b; He et al. 2022; Wang et al. 2023a,c; Zhang et al. 2023a), direct text-to-video diffusion models generates videos of much lower quality and fidelity than the image counterparts (Rombach et al. 2022; Saharia et al. 2022). To alleviate this issue, factorized text-to-video diffusion models (Li et al. 2023; Singer et al. 2023; Girdhar et al. 2023; Blattmann et al. 2023a; Xing et al. 2023; Zhang et al. 2023b; Zeng et al. 2023; Dai et al. 2023) enhance the performance from two perspectives: (i) factorize the generation into text-to-image and image-to-video synthesis and (ii) internally recollect a large amount of high-quality video data. Make-A-Video (Singer et al. 2023) and I2VGen-XL (Zhang et al. 2023b) replace the text embedding with more informative image embedding as input condition. EMU-Video (Girdhar et al. 2023) and Make-Pixels-Dance (Zeng et al. 2023) produce the first frame with the state-of-the-art image models, following by concatenating them to generate high-quality videos. In contrast, our VideoElvator requires no extra training and faithfully integrates the superior capabilities of text-to-image models into sampling chain.

3 Preliminary

Latent diffusion model. Latent Diffusion Model (LDM) (Rombach et al. 2022) mainly consists of pre-trained autoencoder and U-Net (Ronneberger, Fischer, and Brox 2015). The autoencoder uses an encoder $\mathcal{E}$ to compress an image $\boldsymbol{x}$ into latent code $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$ and a decoder $\mathcal{D}$ to reconstruct it. The U-Net is trained to learn the latent distribution $\boldsymbol{z}_0 \sim p_{data}(\boldsymbol{z}_0)$ within the DDPM framework (Ho, Jain, and Abbeel 2020). LDM performs a forward diffusion process by adding Gaussian noise:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{\alpha_t}\boldsymbol{z}_0, \sqrt{1-\alpha_t}\mathbf{I}), \tag{1}$$

where T is the number of diffusion timesteps and $\{\alpha_t\}_{t=1}^T$ control the noise schedules. LDM learns to reverse the above diffusion process through predicting less noisy $\boldsymbol{z}_{t-1}$:

$$p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{z}_{t-1}; \mu_\theta(\boldsymbol{z}_t, t), \Sigma_\theta(\boldsymbol{z}_t, t)), \tag{2}$$

μ_θ and Σ_θ are the mean and variance parameterized with learnable θ. Using Stable Diffusion as base model, T2V and personalized T2I usually freeze the autoencoder during finetuning, so they have shared latent space.

DDIM sampling and inversion. According to Eqn. 1, one can directly predict the clean latent $\boldsymbol{z}_{t\to 0}$ with the noise latent $\boldsymbol{z}_t$ as following:

$$\boldsymbol{z}_{t\to 0} = \frac{\boldsymbol{z}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{z}_t, t)}{\sqrt{\alpha_t}}, \tag{3}$$

where $\epsilon_\theta(\cdot,\cdot)$ denotes "ϵ-prediction" diffusion model. When producing new samples from a random noise $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, DDIM sampling (Song, Meng, and Ermon 2021) denoises $\boldsymbol{z}_t$ to $\boldsymbol{z}_{t-1}$ of previous timestep:

$$\boldsymbol{z}_{t-1} = \sqrt{\alpha_{t-1}}\boldsymbol{z}_{t\to 0} + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta(\boldsymbol{z}_t, t), \tag{4}$$

Given a clean latent $\boldsymbol{z}_0$, DDIM inversion deterministically inverts it into noise latent $\boldsymbol{z}_T$:

$$\boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{z}_{(t-1)\to 0} + \sqrt{1-\alpha_t} \cdot \epsilon_\theta(\boldsymbol{z}_{t-1}, t), \tag{5}$$

The inverted $\boldsymbol{z}_T$ can reconstruct $\boldsymbol{z}_0$ by iterating Eqn. 4. Thanks to its capability in keeping structure, DDIM inversion has been widely used in image and video editing tasks (Hertz et al. 2022; Wu et al. 2023a).

Noise schedules and latent distributions. Noise schedules are parameterized by $\{\alpha_t\}_{t=1}^T$ and define the noise scales at different timesteps (Ho, Jain, and Abbeel 2020; Song et al. 2021). In general, T2V and T2I adopt different noise schedules during training, *i.e*, $\{\alpha_t^V\}_{t=1}^T \neq \{\alpha_t^I\}_{t=1}^T$. Therefore, their noise latent codes belong to different noise distributions, and cannot be used as input for each other.

4 VideoElevator

We introduce VideoElevator to enhance the performance of T2V by integrating T2I into sampling chain, which consists of temporal motion refining (in Sec. 4.1) and spatial quality elevating (in Sec. 4.2). Specifically, VideoElevator reformulates the sampling step at timestep t as follows: taking

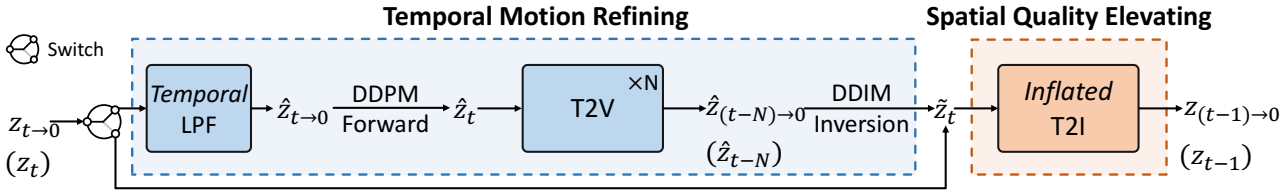

Figure 3: Overview of VideoElevator. Temporal motion refining uses temporal Low-Pass frequency Filter (LPF) to reduce flickers and T2V-based SDEdit to add fine-grained motion, and then inverts the latent to $\tilde{z}_t$ with DDIM inversion. Spatial quality elevating harnesses inflated T2I to directly transition $\tilde{z}_t$ to z_{t-1}, where self-attention of T2I is inflated into cross-frame attention. To ensure interaction between T2V and T2I, noise latents are equally projected to clean latents with Eqn. 3.

noise latent z_t as input, temporal motion refining encapsulates T2V to enhance temporal consistency and inverts noise latent to $\tilde{z}_t$, so that spatial quality elevating directly harnesses T2I to transition it to higher-quality z_{t-1}.

Since T2I and T2V are usually trained in different noise schedules, their noise latent at timestep t may belong to different distributions and cannot be fed into another model. To ensure their interaction, we first project the output z_t of T2I and $\hat{z}_{t-N}$ of T2V into clean latent $z_{t\to 0}$ and $\hat{z}_{(t-N)\to 0}$ with Eqn. 3. Then, we forward clean latent to the noise distributions required by T2V and T2I, but adopt DDPM forward and DDIM inversion respectively. Benefiting from it, VideoElevator supports the combination of various T2V and T2I, as long as their clean latent distributions are shared (*i.e*, same autoencoder).

4.1 Temporal motion refining

Temporal motion refining mainly involves in two aspects: (i) enhance temporal consistency of video latent $z_{t\to 0}$ using T2V generative priors, and (ii) convert it to the noise distribution required by T2I. Firstly, the video decoded from $z_{t\to 0}$ lacks reasonable motion and contains visible flickers. With the power of T2V generative priors (Meng et al. 2021), it is possible to generate a video with natural motion, but challenging to remove its visible flickers (refer to Fig. 6 (a)). Therefore, before using T2V generative priors, we first transform $z_{t\to 0}$ into frequency domain to reduce its high-frequency flickers. Secondly, to preserve the motion integrity in (ii), we deterministically invert video latent into the corresponding noise distribution of T2I.

Specifically, in Fig. 3, we first apply low-pass frequency filter $\texttt{LPF}(\cdot)$ in $z_{t\to 0}$ along the *temporal* dimension, which computes more stable $\hat{z}_{t\to 0}$:

$$\mathcal{F}(z_{t\to 0}) = \texttt{FFT}_{Temp}(z_{t\to 0}), \tag{6}$$

$$\hat{\mathcal{F}}(z_{t\to 0}) = \mathcal{F}(z_{t\to 0}) \odot \mathcal{G}, \tag{7}$$

$$\hat{z}_{t\to 0} = \texttt{IFFT}_{Temp}(\hat{\mathcal{F}}(z_{t\to 0})), \tag{8}$$

where $\texttt{FFT}_{Temp}(\cdot)$ is the fast fourier transformation to map $z_{t\to 0}$ to the frequency domain, while $\texttt{IFFT}_{Temp}(\cdot)$ denotes inverse fast fourier transformation to map it back to the temporal domain. $\mathcal{G}$ represents Gaussian low-pass filter mask to diminish high-frequency flickers. Compared to *spatial-temporal* LPF (Wu et al. 2023b), our *temporal* LPF not only effectively stabilizes the video, but also has less negative impact on spatial quality, *e.g*, more realistic details in Fig. 6.

Albeit there are less high-frequency flickers in $\hat{z}_{t\to 0}$, it is insufficient to achieve vivid and fine-grained motion without T2V generative priors. Inspired by (Meng et al. 2021), we adopt T2V-based SDEdit to portray natural motion to $\hat{z}_{t\to 0}$. In particular, with T2V noise schedule $\{\alpha_t^V\}_{t=1}^T$, we perform DDPM forward process in $\hat{z}_{t\to 0}$ to calculate $\hat{z}_t$ (with Eqn. 1), and then iterates the T2V denoising process N times to achieve desirable motion (with Eqn. 4), *i.e*, $\hat{z}_{t-N}$ or $\hat{z}_{(t-N)\to 0}$. Notably, when N is very small (*e.g*, $N = 1$), the synthesized video only contains coarse-grained motion, so we set N to $8 \sim 10$ to add fine-grained one (refer to Appendix B).

Finally, when inverting $\hat{z}_{(t-N)\to 0}$ into the input noise distribution of T2I, we deterministically forward it to keep its motion. Using the unconditional guidance (Ho and Salimans 2022) (*i.e*, null condition $\emptyset$), clean latent $\hat{z}_{(t-N)\to 0}$ is inverted from timestep 0 to timestep t:

$$\tilde{z}_t = \texttt{Inversion}(\hat{z}_{(t-N)\to 0}, 0, t) \tag{9}$$

We use the inflated T2I model to perform DDIM inversion (Song, Meng, and Ermon 2021) to ensure frame consistency. In contrast, DDPM-based strategies potentially impair the motion integrity of synthesized videos, *e.g*, leading to all frames being identical or inconsistent in Fig. 7.

Empirically, applying temporal motion refining in just a few timesteps (*i.e*, $4 \sim 5$ steps) can ensure temporal consistency (refer to Appendix B). To improve sampling efficiency, we perform both temporal motion refining and spatial quality elevating in selected timesteps, while performing T2I denoising only in other timesteps.

4.2 Spatial quality elevating

Given stablized latent $\tilde{z}_t$ from temporal motion refining, spatial quality elevating leverages T2I to directly add high-quality details. However, individually denoising all frames with conventional T2I will lead to visible inconsistency in appearance. Motivated by previous works (Wu et al. 2023a; Khachatryan et al. 2023), we inflate T2I along the temporal axis so that all frames share the same content.

Particularly, we extend the U-Net of T2I along the temporal dimension, including convolution and self-attention layers. The 2D convolution layers are converted to 3D counterpart by replacing 3×3 kernels with $1 \times 3 \times 3$ kernels. The self-attention layers are extended to first-only cross-frame

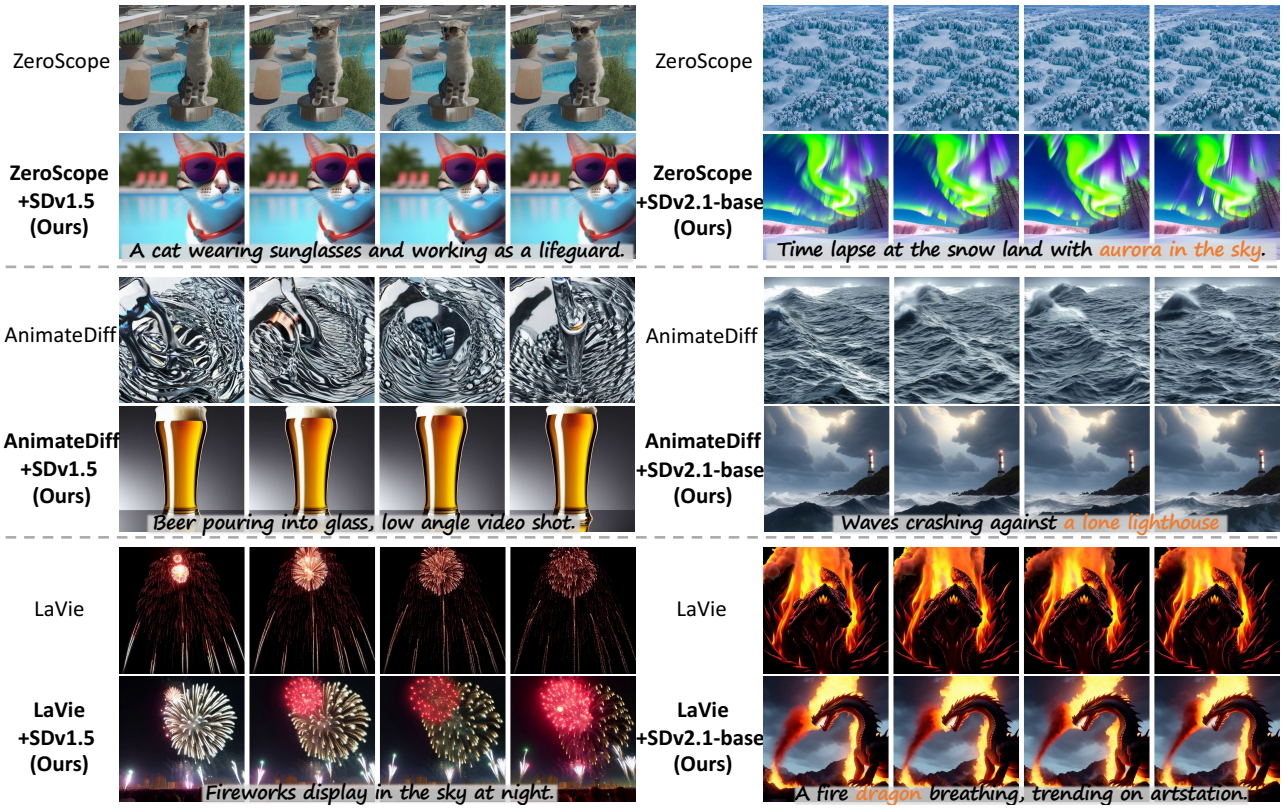

Figure 4: Qualitative results enhanced with foundational T2I. VideoElevator enhances T2V baselines with StableDiffusion V1.5 or V2.1-base in frame quality and text alignment.

attention layers by adding inter-frame interaction. For first-only cross-frame interaction, frame latents $\tilde{\boldsymbol{z}}_t = \{\tilde{\boldsymbol{z}}_t[i]\}_{i=0}^{F-1}$ (F is the number of frames) are updated as:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}) \cdot \boldsymbol{V}, \quad (10)$$

where $\boldsymbol{Q} =^Q \tilde{\boldsymbol{z}}_t[i]$, $\boldsymbol{K} =^K \tilde{\boldsymbol{z}}_t[0]$, and $\boldsymbol{V} =^V \tilde{\boldsymbol{z}}_t[0]$. Afterwards, we directly utilize *inflated* T2I to add photo-realistic details to $\tilde{\boldsymbol{z}}_t$, transitioning it to less noisy latent $\boldsymbol{z}_{t-1}$ at timestep $t-1$:

$$\boldsymbol{z}_{t-1} = \sqrt{\alpha^I_{t-1}} \cdot \tilde{\boldsymbol{z}}_{t\to 0} + \sqrt{1-\alpha^I_{t-1}} \cdot \epsilon^I_\phi(\tilde{\boldsymbol{z}}_t, y, t) + \sigma^I_t \epsilon_t. \quad (11)$$

where ϵ_t is random Gaussian noise and σ^I_t controls its scale.

At each sampling step, spatial quality elevating employs inflated T2I to provide realistic details to video latent. Compared to T2V baselines, VideoElevator produces videos whose frames are closer to T2I generated images, from the perspectives of text alignment, quality, and aesthetic styles.

5 Experiments

5.1 Experimental settings

Evaluation benchmarks. We evaluate VideoElevator and other baselines in two benchmarks: (i) *VBench* (Huang et al. 2023b) dataset that involves in a variety of content categories and contains 800 prompts; (ii) *VideoCreation* dataset, which unifies creative prompts datasets of Make-A-Video (Singer et al. 2023) and VideoLDM (Blattmann et al. 2023b) and consists of 100 prompts in total.

Automated metrics. We utilize five metrics to comprehensively evaluate the performance of T2V: (i) *Frame consistency* (FC) (Zhang et al. 2023c) that calculates the mean CLIP score between each pair of adjacent frames; (ii) *Prompt consistency* (PC) that computes the mean CLIP score between text prompt and all frames; (iii) *Frame quality* (FQ) based on CLIP-IQA (Wang, Chan, and Loy 2023) to quantify the perceived quality, *e.g*, distortions; (iv) *Aesthetic score* (AS) (Huang et al. 2023b) that assesses artistic and beauty value of each frame using the LAION aesthetic predictor; (v) *Domain similarity* (DS) (Guo et al. 2023) that computes the mean CLIP score between all video frames and a reference image.

5.2 Comparisons with T2V baselines

Qualitative comparison. In Fig. 4, VideoElevator visibly improves the frame quality and prompt fidelity of synthesized videos with either Stable Diffusion V1.5 or V2.1-base. In row 1 and 2 of Fig. 4, VideoElevator makes the synthesized videos align better with text prompts than ZeroScope and AnimateDiff, *e.g*, lacking of aurora in the sky and a lone

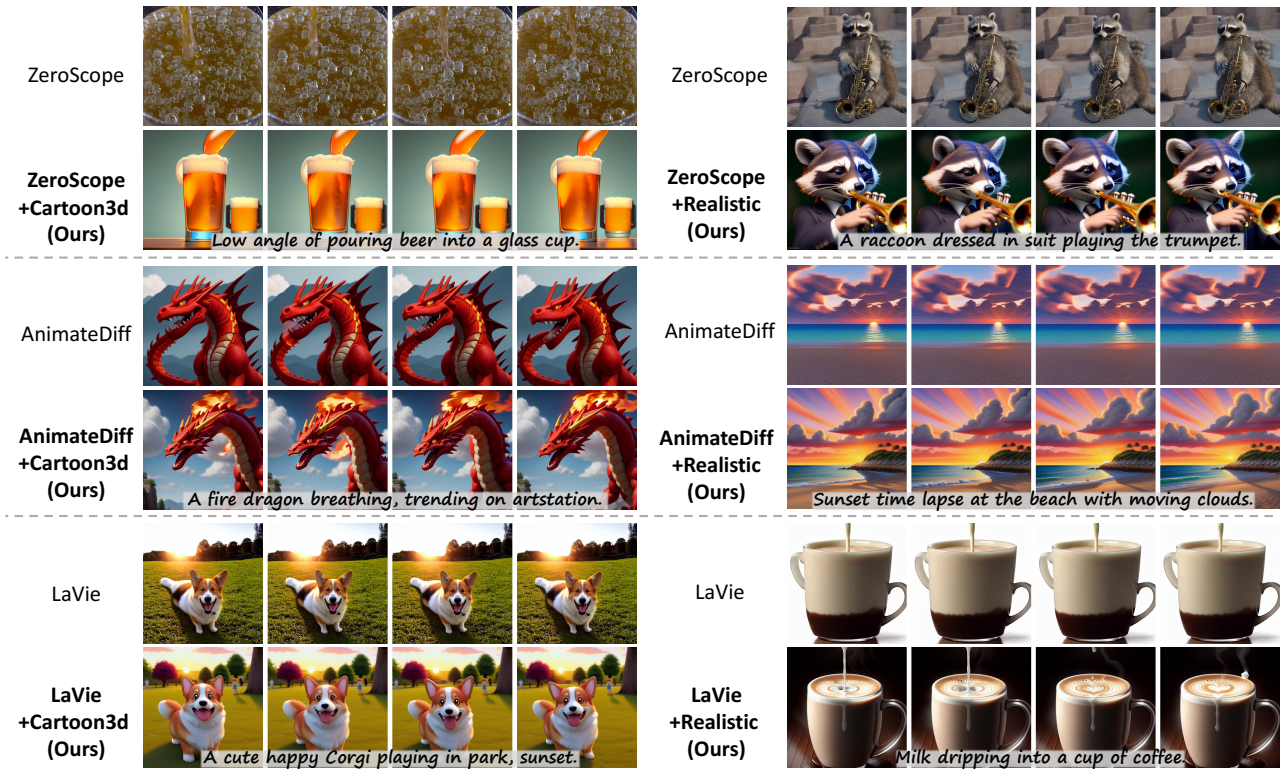

Figure 5: Qualitative results enhanced with personalized T2I. With the power of personalized T2I, VideoElevator enables ZeroScope and LaVie to produce various styles of high-quality videos. Compared to personalized AnimateDiff, VideoElevator captures more faithful styles and photo-realistic details from personalized T2I.

lighthouse. Additionally, in row 3, the videos of VideoElevator have more photo-realistic details than that of LaVie, *e.g*, more colorful fireworks.

From Fig. 5, VideoElevator supports to inherit diverse styles from personalized T2I more faithfully than the alternative competitors. In contrast, personalized AnimateDiff (Radford et al. 2021) captures less-fidelity styles than ours, since its temporal layers suffer from low-quality contents in training videos. ZeroScope (Wang et al. 2023a) and LaVie (Wang et al. 2023c) are not compatible with personalized T2I, as they largely modify the parameters and feature space of initialized T2I.

Quantitative comparison. In row 1 and 2 of Table 1, existing T2V lag far behind foundational T2I in prompt consistency, frame quality, and aesthetic score. With the help of VideoElevator, all T2V baselines are significantly improved in aesthetic score and frame quality, with slightly better frame consistency.

VideoElevator also effectively integrates high-quality T2I into existing T2V, especially in increasing their aesthetic scores. In row 4 and 5 of Table 5, VideoElevator is 0.043 higher than personalized AnimateDiff in terms of domain consistency, *i.e*, capturing higher-fidelity styles, which is consistent to qualitative results.

Human evaluation. We perform human evaluation in VideoCreation dataset. We provide each rater a text prompt and two generated videos from different versions of a baseline (in random order). Then, they are asked to select the better video for each of three perspectives: (i) temporal consistency, (ii) text alignment, and (iii) frame quality. Each sample is evaluated by five raters, and we take a majority vote as the final selection. Table 3 summarizes the voting results of raters. The raters strongly favor the videos from enhanced baselines rather than their base version. In specific, VideoEl-

Method	FC	PC	AS	FQ
SDv1.5	\	0.264	0.633	0.758
SDv2.1-base	\	0.263	0.646	0.714
ZeroScope	0.983	0.245	0.561	0.517
+SDv1.5 (**Ours**)	<u>0.987</u>	**0.252**	<u>0.603</u>	<u>0.576</u>
+SDv2.1-base (**Ours**)	**0.989**	<u>0.248</u>	**0.618**	**0.593**
AnimateDiff	0.983	0.248	0.582	0.561
+SDv1.5 (**Ours**)	<u>0.984</u>	**0.252**	<u>0.610</u>	<u>0.617</u>
+SDv2.1-base (**Ours**)	**0.985**	**0.252**	**0.621**	**0.621**
LaVie	0.991	0.255	0.607	0.681
+SDv1.5 (**Ours**)	**0.993**	**0.259**	<u>0.627</u>	**0.706**
+SDv2.1-base (**Ours**)	<u>0.992</u>	<u>0.256</u>	**0.632**	<u>0.702</u>

Table 1: Quantitative results using foundational T2I on VBench benchmark. The best and second best results are bolded and underlined.

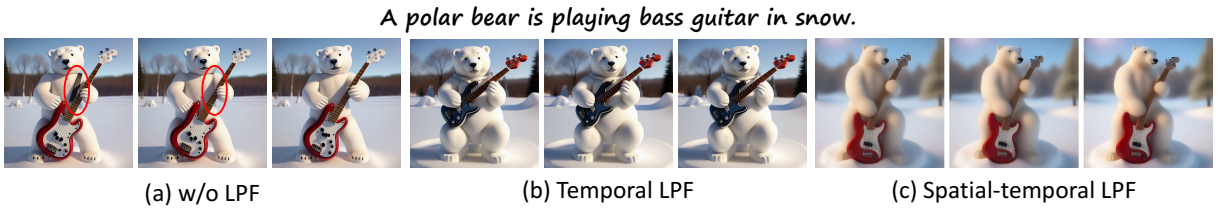

Figure 6: Qualitative ablation on low-pass frequency filter (LPF). (a) Not using LPF leads to incoherence in appearance (highlighted in red circle). (b) Using temporal LPF visibly improves temporal consistency without hurting frame quality. (c) Adding both spatial and temporal LPF makes video frames blurry.

Figure 7: Quantitative ablation on inversion strategies. We visualize synthesized frames and x-t slice of pixels in red line.

Method	FC	PC	AS	FQ	DS
SD-LoRA	\	0.288	0.702	0.790	1.000
ZeroScope	0.981	0.276	0.578	0.544	0.700
+SD-LoRA (**Ours**)	**0.991**	**0.281**	**0.671**	**0.684**	**0.871**
AnimateDiff	0.993	0.275	0.667	0.670	0.830
+SD-LoRA (**Ours**)	**0.993**	**0.280**	**0.677**	**0.689**	**0.873**
LaVie	0.991	0.281	0.628	0.728	0.773
+SD-LoRA (**Ours**)	**0.992**	**0.285**	**0.680**	**0.729**	**0.852**

Table 2: Quantitative results using personalized T2I on VideoCreation benchmark. The best results are bolded.

Method comparison	Temporal cons.	Text align.	Frame qual.
Ours vs. ZeroScope	65%	81%	88%
Ours vs. AnimateDiff	62%	75%	86%
Ours vs. LaVie	61%	72%	81%

Table 3: User preference study.

evator manages to boost their performance in terms of text alignment and frame quality.

5.3 Ablation studies

Effect of low-pass frequency filter. We probe its effect under three settings: (i) without LPF, (ii) temporal LPF, and (iii) spatial-temporal LPF (Wu et al. 2023b). Their experimental results are shown in Fig. 6. In Fig. 6 (a), w/o LPF results in appearance incoherency, *e.g*, disappearing black tape. In contrast, adding LPF along temporal dimension noticeably improves frame consistency. However, applying LPF along both spatial and temporal dimensions visibly degrades frame quality and aesthetic score, making frames blurry and less detailed.

Effect of inversion strategies. We investigate three inversion strategies in Fig. 7: (i) perturb all frames with random noise (Eqn. 1), (ii) perturb all frames with same noise (Eqn. 1), and (iii) DDIM inversion. Compared to adding random noise, using DDIM inversion achieves better frame consistency and more continuous x-t slice. In Fig. 7 (a), albeit perturbing with same noise obtains higher frame consistency, it results in all frames becoming identical.

6 Conclusion

We introduce VideoElevator, a *play-and-plug* approach that boosts the performance of T2V baselines with versatile T2I. VideoElevator explicitly decomposes each step into temporal motion refining and spatial quality elevating. Given noise latent at timestep t, temporal motion refining leverages a low-pass frequency filter to reduce its flickers, while applying T2V-based SDEdit to portray realistic motion. Spatial quality elevating inflates the self-attention of T2I along the temporal axis, and directly harnesses inflated T2I to predict less noisy latent. Extensive experiments demonstrate the effectiveness of VideoElevator under the combination of various T2V and T2I models. Even with foundational T2I, VideoElevator significantly improves T2V baselines in terms of text alignment, frame quality, and aesthetic score. When integrating personalized T2I, it enables T2V baselines to faithfully produce various styles of videos.

Acknowledgements

This work was supported by the National Key R&D Program of China (2021YFF0900500) and the National Natural Science Foundation of China (NSFC) under grants 62441202.

References

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*.

Bar-Tal, O.; Chefer, H.; Tov, O.; Herrmann, C.; Paiss, R.; Zada, S.; Ephrat, A.; Hur, J.; Li, Y.; Michaeli, T.; et al. 2024. Lumiere: A Space-Time Diffusion Model for Video Generation. *arXiv preprint arXiv:2401.12945*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023b. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *CVPR*.

Cai, Y.; Wei, Y.; Ji, Z.; Bai, J.; Han, H.; and Zuo, W. 2023. Decoupled Textual Embeddings for Customized Image Generation. *arXiv preprint arXiv:2312.11826*.

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; et al. 2023. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. *arXiv preprint arXiv:2310.19512*.

Dai, Z.; Zhang, Z.; Yao, Y.; Qiu, B.; Zhu, S.; Qin, L.; and Wang, W. 2023. Fine-Grained Open Domain Image Animation with Motion Guidance. *arXiv preprint arXiv:2311.12886*.

Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; and Germanidis, A. 2023. Structure and content-guided video synthesis with diffusion models. In *ICCV*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *ICLR*.

Ge, S.; Nah, S.; Liu, G.; Poon, T.; Tao, A.; Catanzaro, B.; Jacobs, D.; Huang, J.-B.; Liu, M.-Y.; and Balaji, Y. 2023. Preserve your own correlation: A noise prior for video diffusion models. In *ICCV*.

Girdhar, R.; Singh, M.; Brown, A.; Duval, Q.; Azadi, S.; Rambhatla, S. S.; Shah, A.; Yin, X.; Parikh, D.; and Misra, I. 2023. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*.

Guo, Y.; Yang, C.; Rao, A.; Wang, Y.; Qiao, Y.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent Video Diffusion Models for High-Fidelity Video Generation with Arbitrary Lengths. *arXiv preprint arXiv:2211.13221*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, T.; Zeng, Y.; Zhang, Z.; Xu, W.; Xu, H.; Xu, S.; Lau, R. W.; and Zuo, W. 2023a. Dreamcontrol: Control-based text-to-3d generation with 3d self-prior. *arXiv preprint arXiv:2312.06439*.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2023b. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. *arXiv preprint arXiv:2303.13439*.

Li, Z.; Tucker, R.; Snavely, N.; and Holynski, A. 2023. Generative image dynamics. *arXiv preprint arXiv:2309.07906*.

Liang, J.; Fan, Y.; Zhang, K.; Timofte, R.; Van Gool, L.; and Ranjan, R. 2023. MoVideo: Motion-Aware Video Generation with Diffusion Models. *arXiv preprint arXiv:2311.11325*.

Lin, J.; Zhang, Z.; Wei, Y.; Ren, D.; Jiang, D.; and Zuo, W. 2023. Improving image restoration through removing degradations in textual representations. *arXiv preprint arXiv:2312.17334*.

Lv, Z.; Wei, Y.; Zuo, W.; and Wong, K.-Y. K. 2024. PLACE: Adaptive Layout-Semantic Fusion for Semantic Image Synthesis. *IEEE Conference on Computer Vision and Pattern Recognition*.

Ma, Y.; Cun, X.; He, Y.; Qi, C.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023a. MagicStick: Controllable Video Editing via Control Handle Transformations. *arXiv preprint arXiv:2312.03047*.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023b. Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos. *arXiv preprint arXiv:2304.01186*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Ni, M.; Zhang, Y.; Feng, K.; Li, X.; Guo, Y.; and Zuo, W. 2023. Ref-diff: Zero-shot referring image segmentation with generative models. *arXiv preprint arXiv:2308.16777*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *ICML*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding.

Sauer, A.; Karras, T.; Laine, S.; Geiger, A.; and Aila, T. 2023. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. *arXiv preprint arXiv:2301.09515*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2023. Make-a-video: Text-to-video generation without text-video data. In *ICLR*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising diffusion implicit models. In *ICLR*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2023. Diffusion Model Alignment Using Direct Preference Optimization. *arXiv preprint arXiv:2311.12908*.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2555–2563.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023a. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, X.; Yuan, H.; Zhang, S.; Chen, D.; Wang, J.; Zhang, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023b. VideoComposer: Compositional Video Synthesis with Motion Controllability. *arXiv preprint arXiv:2306.02018*.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2023c. LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models. *arXiv preprint arXiv:2309.15103*.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023a. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7623–7633.

Wu, T.; Si, C.; Jiang, Y.; Huang, Z.; and Liu, Z. 2023b. FreeInit: Bridging Initialization Gap in Video Diffusion Models. *arXiv preprint arXiv:2312.07537*.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Wang, X.; Wong, T.-T.; and Shan, Y. 2023. DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors. *arXiv preprint arXiv:2310.12190*.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.

Zeng, Y.; Wei, G.; Zheng, J.; Zou, J.; Wei, Y.; Zhang, Y.; and Li, H. 2023. Make pixels dance: High-dynamic video generation. *arXiv preprint arXiv:2311.10982*.

Zhang, D. J.; Wu, J. Z.; Liu, J.-W.; Zhao, R.; Ran, L.; Gu, Y.; Gao, D.; and Shou, M. Z. 2023a. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; and Zhou, J. 2023b. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023c. ControlVideo: Training-free Controllable Text-to-Video Generation. *arXiv preprint arXiv:2305.13077*.