International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02295-1

LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models

Yaohui Wang[1] · Xinyuan Chen[1] · Xin Ma[1,4] · Shangchen Zhou[2] · Ziqi Huang[2] · Yi Wang[1] · Ceyuan Yang[1] · Yinan He[1] · Jiashuo Yu[1] · Peiqing Yang[2] · Yuwei Guo[1,3] · Tianxing Wu[2] · Chenyang Si[2] · Yuming Jiang[2] · Cunjian Chen[4] · Chen Change Loy[2] · Bo Dai[1] · Dahua Lin[1,3] · Yu Qiao[1] · Ziwei Liu[2]

Received: 29 March 2024 / Accepted: 28 October 2024

Abstract

This work aims to learn a high-quality text-to-video (T2V) generative model by leveraging a pre-trained text-to-image (T2I) model as a basis. It is a highly desirable yet challenging task to simultaneously (a) accomplish the synthesis of visually realistic and temporally coherent videos while (b) preserving the strong creative generation nature of the pre-trained T2I model. To this end, we propose **LaVie**, an integrated video generation framework that operates on cascaded video latent diffusion models, comprising a base T2V model, a temporal interpolation model, and a video super-resolution model. Our key insights are two-fold: (1) We reveal that the incorporation of simple temporal self-attentions, coupled with rotary positional encoding, adequately captures the temporal correlations inherent in video data. (2) Additionally, we validate that the process of joint image-video fine-tuning plays a pivotal role in producing high-quality and creative outcomes. To enhance the performance of LaVie, we contribute a comprehensive and diverse video dataset named **Vimeo25M**, consisting of 25 million text-video pairs that prioritize quality, diversity, and aesthetic appeal. Extensive experiments demonstrate that LaVie achieves state-of-the-art performance both quantitatively and qualitatively. Furthermore, we showcase the versatility of pre-trained LaVie models in various long video generation and personalized video synthesis applications. Project page: https://github.com/Vchitect/LaVie/.

Keywords Video generation · Diffusion models · Generative modeling

1 Introduction

With the remarkable breakthroughs achieved by Diffusion Models (DMs) (Ho et al., 2020; Song et al., 2021a, b) in image synthesis, the generation of photorealistic images from text descriptions (T2I) (Ramesh et al., 2021, 2022; Saharia et al., 2022; Balaji et al., 2022; Rombach et al., 2022) has taken center stage, finding applications in various image processing domain such as image outpainting (Ramesh et al., 2022), editing (Zhang & Agrawala, 2023; Mokady et al., 2022; Parmar et al., 2023; Huang et al., 2023) and enhancement (Saharia et al., 2022; Wang et al., 2023). Building upon the successes of T2I models, there has been a growing interest in extending these techniques to the synthesis of videos controlled by text inputs (T2V) (Singer et al., 2023; Ho et al., 2022; Blattmann et al., 2023; Zhou et al., 2022; He et al., 2022), driven by their potential applications in domains such as filmmaking, video games, and artistic creation.

Communicated by Shengfeng He.

Yaohui Wang, Xinyuan Chen and Xin Ma have contributed equally to this work.

✉ Yaohui Wang
wangyaohui@pjlab.org.cn

✉ Dahua Lin
dhlin@ie.cuhk.edu.hk

✉ Yu Qiao
qiaoyu@pjlab.org.cn

✉ Ziwei Liu
ziwei.liu@ntu.edu.sg

[1] Shanghai Artificial Intelligence Laboratory, Shanghai, China

[2] S-Lab, Nanyang Technological University, Singapore, Singapore

[3] The Chinese University of Hong Kong, Hong Kong SAR, China

[4] Monash University, Melbourne, Australia

Published online: 23 December 2024

However, training an entire T2V system from scratch (Ho et al., 2022) poses significant challenges as it requires extensive computational resources to optimize the entire network for learning spatio-temporal joint distribution. An alternative approach (Singer et al., 2023; Blattmann et al., 2023; Zhou et al., 2022; He et al., 2022) leverages the prior spatial knowledge from pre-trained T2I models for faster convergence to adapt video data, which aims to expedite the training process and efficiently achieve high-quality results. However, in practice, finding the right balance among video quality, training cost, and model compositionality still remains challenging as it required careful design of model architecture, training strategies and the collection of high-quality text-video datasets.

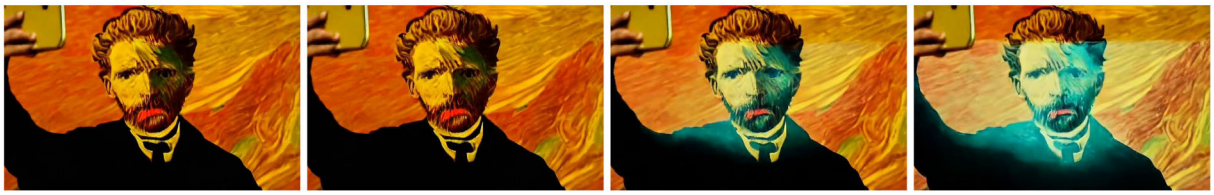
To this end, we introduce **LaVie**, an integrated video generation framework (with a total number of 3B parameters) that operates on cascaded video latent diffusion models. LaVie is a text-to-video foundation model built based on a pre-trained T2I model (i.e. Stable Diffusion (Rombach et al., 2022)), aiming to synthesize visually realistic and temporally coherent videos while preserving the strong creative generation nature of the pre-trained T2I model. Our key insights are two-fold: 1) simple temporal self-attention coupled with RoPE (Su et al., 2021) adequately captures temporal correlations inherent in video data. More complex architectural design only results in marginal visual improvements to the generated outcomes. 2) Joint image-video fine-tuning plays a key role in producing high-quality and creative results. Directly fine-tuning on video dataset severely hampers the concept-mixing ability of the model, leading to *catastrophic forgetting* and the gradual vanishing of learned prior knowledge. Moreover, joint image-video fine-tuning facilitates large-scale knowledge transferring from images to videos, encompassing scenes, styles, and characters. In addition, we found that current publicly available text-video dataset WebVid10M (Bain et al., 2021), is insufficient to support T2V task due to its low resolution and watermark-centered videos. Therefore, to enhance the performance of LaVie, we introduce a novel text-video dataset **Vimeo25M** which consists of 25 million high-resolution videos ($>$ 720p) with text descriptions. Our experiments demonstrate that training on Vimeo25M substantially boosts the performance of LaVie and empowers it to produce superior results in terms of quality, diversity, and aesthetic appeal (see Fig. 1 and Fig. 2).

2 Related Work

Unconditional video generation endeavors to generate videos by comprehensively learning the underlying distribution of the training dataset. Previous works have leveraged various types of deep generative models, including GANs (Goodfellow et al., 2014; Radford et al., 2015; Brock et al., 2019; Karras et al., 2019, 2020; Vondrick et al., 2016; Saito et al., 2017; Tulyakov et al., 2018; WANG et al., 2020; Wang et al., 2020; Wang, 2021; Wang et al., 2021; Clark et al., 2019; Brooks et al., 2022; Chen et al., 2020; Yu et al., 2022; Skorokhodov et al., 2022; Tian et al., 2021; Zhang et al., 2023), VAEs (Kingma & Welling, 2014; Li & Mandt, 2018; Bhagat et al., 2020; Xie et al., 2020), and VQ-based models (Van Den Oord et al., 2017; Esser et al., 2021; Yan et al., 2021; Ge et al., 2022; Jiang et al., 2023). Recently, a notable advancement in video generation has been observed with the emergence of Diffusion Models (DMs) (Ho et al., 2020; Song et al., 2021a; Nichol & Dhariwal, 2021), which have demonstrated remarkable progress in image synthesis (Ramesh et al., 2021, 2022; Rombach et al., 2022). Building upon this success, several recent works (Ho et al., 2022; He et al., 2022; Wang et al., 2023) have explored the application of DMs for video generation. These works showcase the promising capability of DMs to model complex video distributions by integrating spatio-temporal operations into image-based models, surpassing previous approaches in terms of video quality. However, learning the entire distribution of video datasets in an unconditional manner remains highly challenging. The entanglement of spatial and temporal content poses difficulties, making it still arduous to obtain satisfactory results.

Text-to-video generation, as a form of conditional video generation, focuses on the synthesis of high-quality videos using text descriptions as conditioning inputs. Existing approaches primarily extend text-to-image models by incorporating temporal modules, such as temporal convolutions and temporal attention, to establish temporal correlations between video frames.

Notably, previous works (Singer et al., 2023; Ho et al., 2022; Ge et al., 2023; Blattmann et al., 2023; Zhou et al., 2022; He et al., 2022; Chen et al., 2024; Zhang et al., 2023; Blattmann et al., 2023) developed UNet-based text-to-video diffusion models based on pre-trained text-to-image models (Ramesh et al., 2021, 2022; Balaji et al., 2022; Rombach et al., 2022). These approaches follow the standard cascaded pipeline design and train the entire system on large-scale video datasets. Other works (Guo et al., 2023; Zhang et al., 2024) focus on constructing the personalized text-to-video model to partially train temporal modules which can be adapted to various different personalized text-to-image models for personalized video generation. Recently, DiT (Peebles & Xie, 2023) has started to replace UNet to become the novel architecture for text-to-image diffusion models (Chen et al., 2024; Esser et al., 2024). Several latest works (Brooks et al., 2024; Ma et al., 2024; Gupta et al., 2023; Lu et al., 2023; Yang et al., 2024) extend such idea on video generation and show promising results on large-scale text-to-video generation.



Cinematic shot of Van Gogh's selfie, Van Gogh style.

A happy panda in space suit walking in the space.

The Bund, Shanghai, with the ship moving on the river, oil painting.

A super cool giant robot in Cyberpunk city, artstation.

A fantasy landscape, trending on artstation, 4k, high resolution.

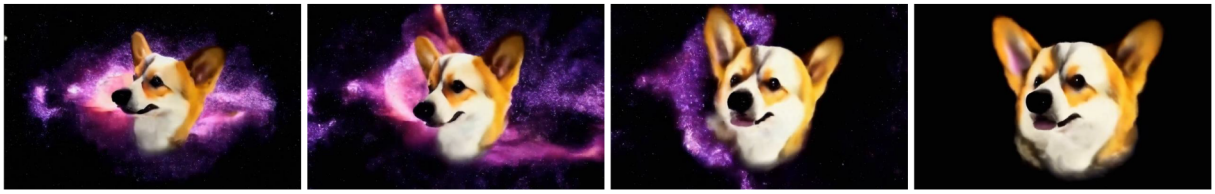
Fig. 1 Text-to-video samples. LaVie is able to synthesize diverse, creative, high-definition videos with photorealistic and temporal coherent content by giving text descriptions

3 Preliminary of Diffusion Models

Diffusion models (DMs) (Ho et al., 2020; Song et al., 2021a, b) aim to learn the underlying data distribution through a combination of two fundamental processes: *diffusion* and *denoising*. Given an input data sample $z \sim p(z)$, the diffusion process introduces random noises to construct a noisy sample $z_t = \alpha_t z + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathrm{I})$. This process is achieved by a Markov chain with T steps, and the noise scheduler is parametrized by the diffusion-time t, characterized by α_t and σ_t. Notably, the logarithmic signal-to-noise ratio $\lambda_t = log[\alpha^2 t/\sigma^2 t]$ monotonically decreases over time. In the subsequent denoising stage, ϵ-prediction and v-prediction are employed to learn a denoiser function ϵ_θ, which is trained to minimize the mean square error loss by taking the diffused sample z_t as input:

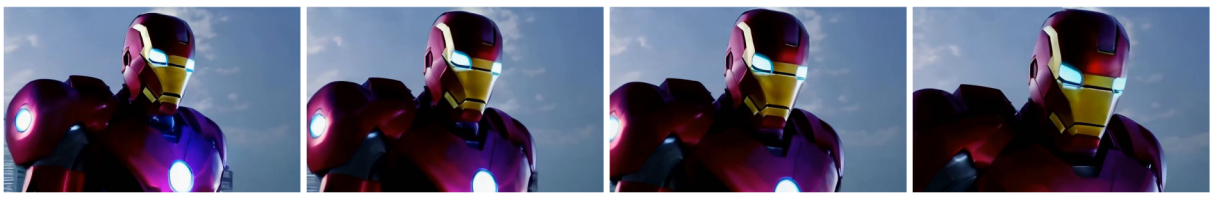
$$\mathbb{E}_{\mathbf{z}\sim p(z),\ \epsilon\sim\mathcal{N}(0,1),\ t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t)\|_2^2\right]. \tag{1}$$

Latent diffusion models (LDMs) (Rombach et al., 2022) utilize a variational autoencoder architecture, wherein the encoder $\mathcal{E}$ is employed to compress the input data into low-dimensional latent codes $\mathcal{E}(z)$. Diverging from previous



A corgi's head depicted as an explosion of a nebula, high quality.

Gwen Stacy reading a book.



Iron Man flying in the sky.

Yoda playing guitar on the stage.

A shark swimming in the ocean.

Fig. 2 Text-to-video samples. LaVie is able to synthesize diverse, creative, high-definition videos with photorealistic and temporal coherent content by giving text descriptions

methods, LDMs conduct the diffusion and denoising processes in the latent space rather than the data space, resulting in substantial reductions in both training and inference time. Following the denoising stage, the final output is decoded as $\mathcal{D}(z_0)$, representing the reconstructed data. The objective of LDMs can be formulated as follows:

$$\mathbb{E}_{\mathbf{z}\sim p(z),\ \epsilon\sim\mathcal{N}(0,1),\ t}\left[\|\epsilon-\epsilon_\theta(\mathcal{E}(\mathbf{z}_t),t)\|_2^2\right]. \tag{2}$$

Our proposed LaVie follows the idea of LDMs to encode each video frames into per frame latent code $\mathcal{E}(z)$. The diffusion process is operated in the latent spatio-temporal distribution space to model latent video distribution.

4 Our Approach

Our proposed framework, LaVie, is a cascaded framework consisting of Video Latent Diffusion Models (V-LDMs) conditioned on text descriptions. The overall architecture of LaVie is depicted in Fig. 3, and it comprises three distinct networks: a Base T2V model responsible for generating

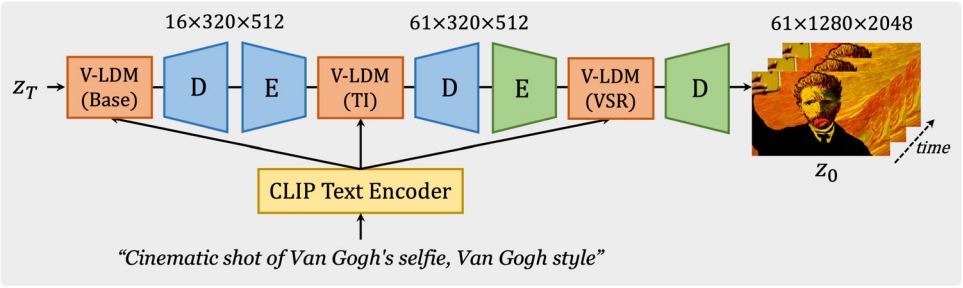


Fig. 3 General pipeline. LaVie consists of three modules: a Base T2V model, a Temporal Interpolation (TI) model, and a Video Super Resolution (VSR) model. It also requires Encoder (E) and Decoder (D) from pretrained VAE. At the inference stage, given a sequence of noise and a text description, the base model aims to generate key frames aligning with the prompt and containing temporal correlation. The temporal interpolation model focuses on producing smoother results and synthesizing richer temporal details. The video super-resolution model enhances the visual quality as well as elevates the spatial resolution even further. Finally, we generate videos at 1280 × 2048 resolution with 61 frames

short, low-resolution key frames, a Temporal Interpolation (TI) model designed to interpolate the short videos and increase the frame rate, and a Video Super Resolution (VSR) model aimed at synthesizing high-definition results from the low-resolution videos. Each of these models is individually trained with text inputs serving as conditioning information. During the inference stage, given a sequence of latent noises and a textual prompt, LaVie is capable of generating a video consisting of 61 frames with a spatial resolution of 1280 × 2048 pixels, utilizing the entire system. In the subsequent sections, we will elaborate on the learning methodology employed in LaVie, as well as the architectural design of the models involved.

4.1 Base T2V Model

Given the video dataset p_{video} and the image dataset p_{image}, we have a T-frame video denoted as $v \in \mathbb{R}^{T\times 3\times H\times W}$, where v follows the distribution p_{video}. Similarly, we have an image denoted as $x \in \mathbb{R}^{3\times H\times W}$, where x follows the distribution p_{image}. As the original LDM is designed as a 2D UNet and can only process image data, we introduce two modifications to model the spatio-temporal distribution. Firstly, for each 2D convolutional layer, we inflate the pre-trained kernel to incorporate an additional temporal dimension, resulting in a pseudo-3D convolutional layer. This inflation process converts any input tensor with the size $B \times C \times H \times W$ to $B \times C \times 1 \times H \times W$ by introducing an extra temporal axis. Secondly, as illustrated in Fig. 4, we extend the original transformer block to a Spatio-Temporal Transformer (ST-Transformer) by including a temporal attention layer after each spatial layer. Furthermore, we incorporate the concept of Rotary Positional Encoding (RoPE) from the recent LLM (Touvron et al., 2023) to integrate the temporal attention layer. Unlike previous methods that introduce an additional Temporal Transformer to model time, our modification directly applies to the transformer block itself, resulting in a simpler yet effective approach. Through various experiments with different designs of the temporal module, such as spatio-temporal attention and temporal causal attention, we observed that increasing the complexity of the temporal module only marginally improved the results while significantly increasing model size and training time. Therefore, we opt to retain the simplest design of the network, generating videos with 16 frames at a resolution of 320×512.

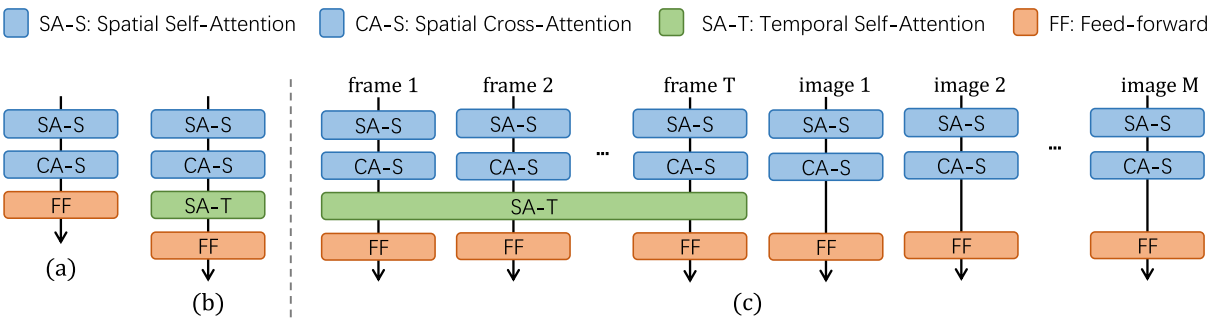


Fig. 4 Spatio-temporal module. We show the Transformer block in Stable Diffusion in (**a**), our proposed ST-Transformer block in (**b**), and our joint image-video training scheme in (**c**)

The primary objective of the base model is to generate high-quality key frames while also preserving diversity and capturing the compositional nature of videos. We aim to enable our model to synthesize videos aligned with creative prompts, such as *"Cinematic shot of Van Gogh's selfie"*. However, we observed that fine-tuning solely on video datasets, even with the initialization from a pre-trained LDM, fails to achieve this goal due to the phenomenon of catastrophic forgetting, where previous knowledge is rapidly forgotten after training for a few epochs. Hence, we apply a joint fine-tuning approach using both image and video data to address this issue. In practise, we concatenate M images along the temporal axis to form a T-frame video and train the entire base model to optimize the objectives of both the Text-to-Image (T2I) and Text-to-Video (T2V) tasks (as shown in Fig. 4 c). Consequently, our training objective consists of two components: a video loss $\mathcal{L}_V$ and an image loss $\mathcal{L}_I$. The overall objective can be formulated as:

$$\mathcal{L} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(\mathcal{E}(\mathbf{v}_t), t, c_V)\|_2^2\right] + \tag{3}$$

$$\alpha * \mathbb{E}\left[\|\epsilon - \epsilon_\theta(\mathcal{E}(\mathbf{x}_t), t, c_I)\|_2^2\right], \tag{4}$$

where c_V and c_I represent the text descriptions for videos and images, respectively, and α is the coefficient used to balance the two losses. By incorporating images into the fine-tuning process, we observe a significant improvement in video quality. Furthermore, as demonstrated in Fig. 1, our approach successfully transfers various concepts from images to videos, including different styles, scenes, and characters. An additional advantage of our method is that, since we do not modify the architecture of LDM and jointly train on both image and video data, the resulting base model is capable of handling both T2I and T2V tasks, thereby showcasing the generalizability of our proposed design.

We use two-stage training strategy to train our base model. Firstly, we pre-train the model on large-scale video-text and image-text datasets towards enabling the model to capture the diverse spatio-temporal concepts. In the second stage, we fine-tune the pre-trained model on a relatively smaller-scale dataset with higher-quality. Similar to Dai et al. (2023), we found using such strategy is able to further improve the generated quality.

4.2 Temporal Interpolation Model

Building upon our base T2V model, we introduce a temporal interpolation network to enhance the smoothness of our generated videos and synthesize richer temporal details (see Fig. 5). We accomplish this by training a diffusion UNet, designed specifically to quadruple the frame rate of the base video. This network takes a 16-frame base video as input and produces an upsampled output consisting of 61 frames. During the training phase, we duplicate the base video frames to match the target frame rate and concatenate them with the noisy high-frame-rate frames. This combined data is fed into the diffusion UNet. We train the UNet using the objective of reconstructing the noise-free high-frame-rate frames, enabling it to learn the process of denoising and generate the interpolated frames. At inference time, the base video frames are concatenated with randomly initialized Gaussian noise. The diffusion UNet gradually removes this noise through the denoising process, resulting in the generation of the 61 interpolated frames. Notably, our approach differs from conventional video frame interpolation methods, as each frame generated through interpolation replaces the corresponding input frame. In other words, every frame in the output is newly synthesized, providing a distinct approach compared to techniques where the input frames remain unchanged during interpolation. Furthermore, our diffusion UNet is conditioned on the text prompt, which serves as additional guidance for the temporal interpolation process, enhancing the overall quality and coherence of the generated videos.

4.3 Video Super Resolution Model

To further enhance visual quality and elevate spatial resolution, we incorporate a video super-resolution (VSR) model into our video generation pipeline. This involves training a LDM upsampler, specifically designed to increase the video resolution to 1280×2048. Similar to the base model described in Sect. 4.1, we leverage a pre-trained diffusion-based image $\times 4$ upscaler as a prior. To adapt the network architecture to process video inputs in 3D, we incorporate an additional temporal dimension, enabling temporal processing within the diffusion UNet. Within this network, we introduce temporal layers, namely temporal attention and a 3D convolutional layer, alongside the existing spatial layers. These temporal layers contribute to enhancing temporal coherence in the generated videos. By concatenating the low-resolution input frames within the latent space, the diffusion UNet takes into account additional text descriptions and noise levels as conditions, which allows for more flexible control over the texture and quality of the enhanced output.

While the spatial layers in the pre-trained upscaler remain fixed, our focus lies in fine-tuning the inserted temporal layers in the V-LDM. Inspired by CNN-based super-resolution networks (Chan et al., 2022a, b; Zhou et al., 2022, 2020; Jiang et al., 2021, 2022), our model undergoes patch-wise training on 320×320 patches. By utilizing the low-resolution video as a strong condition, our upscaler UNet effectively preserves its intrinsic convolutional characteristics. This allows for efficient training on patches while maintaining the capability to process inputs of arbitrary sizes. Through the integration of the VSR model, our LaVie framework generates high-quality

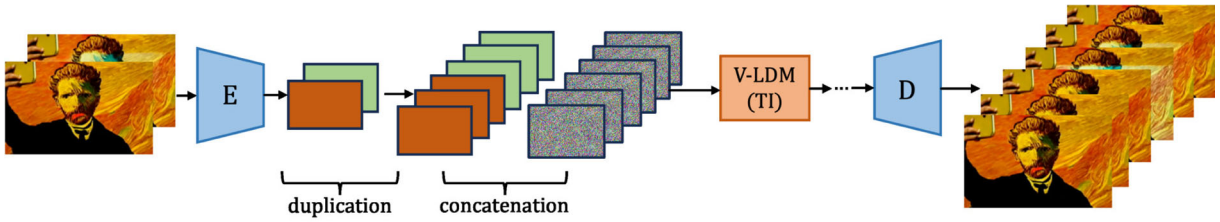


Fig. 5 Inference of temporal interpolation model. At inference time, we duplicate the base video frames to match the target frame rate and concatenate them with randomly initialized Gaussian noise. The diffusion UNet gradually removes this noise through the denoising process, resulting in the generation of the 61 interpolated frames

videos at a 2K resolution (1280×2048), ensuring both visual excellence and temporal consistency in the final output.

5 Experiments

In this section, we present our experimental settings, encompassing datasets and implementation details. Subsequently, we evaluate our method both qualitatively and quantitatively, comparing it to state-of-the-art on the zero-shot text-to-video task. We then conduct an in-depth analysis regarding the efficacy of joint image-video fine-tuning. Next, we showcase two applications of our method: long video generation and personalized video synthesis. Finally, we discuss limitations and potential solutions to improve current approach.

5.1 Datasets

To train our models, we leverage two publicly available datasets, namely Webvid10M (Bain et al., 2021) and Laion5B (Schuhmann et al., 2022). However, we encountered limitations when utilizing WebVid10M for high-definition video generation, specifically regarding video resolution, diversity, and aesthetics. Therefore, we curate a new dataset called Vimeo25M, specifically designed to enhance the quality of text-to-video generation. By applying rigorous filtering criteria based on resolution and aesthetic scores, we obtained a total of 20 million videos and 400 million images for training purposes.

Vimeo25M dataset. A collection of 25 million text-video pairs in high-definition, widescreen, and watermark-free formats. These pairs are automatically generated using Videochat (Li et al., 2023). The original videos are sourced from Vimeo[1] and are classified into ten categories: *Ads and Commercials*, *Animation*, *Branded Content*, *Comedy*, *Documentary*, *Experimental*, *Music*, *Narrative*, *Sports*, and *Travel*. Example videos are shown in Fig. 6. To obtain the dataset, we utilized PySceneDetect[2] for scene detection and segmentation of the primary videos. To ensure the quality of captions, we filtered out captions with less than three words and excluded video segments with fewer than 16 frames. Consequently, we obtained a total of 25 million individual video segments, each representing a single scene. The statistics of the Vimeo25M dataset, including the distribution of video categories, the duration of video segments, and the length of captions, are presented in Fig. 7. The dataset demonstrates a diverse range of categories, with a relatively balanced quantity among the majority of categories. Moreover, most videos in the dataset have captions consisting of approximately 10 words.

We conducted a comparison of the aesthetics score between the Vimeo25M dataset and the WebVid10M dataset. As illustrated in Fig. 8a, approximately 16.89% of the videos in Vimeo25M received a higher aesthetics score (greater than 6), surpassing the 7.22% in WebVid10M. In the score range between 4 and 6, Vimeo25M achieved a percentage of 79.12%, which is also superior to the 72.58% in WebVid10M. Finally, Fig. 8b depicts a comparison of the spatial resolution between the Vimeo25M and WebVid10M datasets. It is evident that the majority of videos in the Vimeo25M dataset possess a higher resolution than those in WebVid10M, thereby ensuring that the generated results exhibit enhanced quality.

Our high-quality dataset contains around 2000 video-text pairs. We filter out data based on aesthetic score, clip score and video-text alignment.

5.2 Implementation Details

The Autoencoder and LDM of Base T2V model is initialized from a pretrained Stable Diffusion 1.4. Prior to training, we preprocess each video to a resolution of 320×512 and train using 16 frames per video clip. Additionally, we concatenate 4 images to each video for joint image-video fine-tuning. To facilitate the fine-tuning process, we employ curriculum

[1] https://vimeo.com

[2] https://github.com/Breakthrough/PySceneDetect

(a) *An aerial view of a large estate.*

(b) *A sunset with clouds in the sky.*

Fig. 6 We show three video examples as well as text descriptions from Vimeo25M dataset

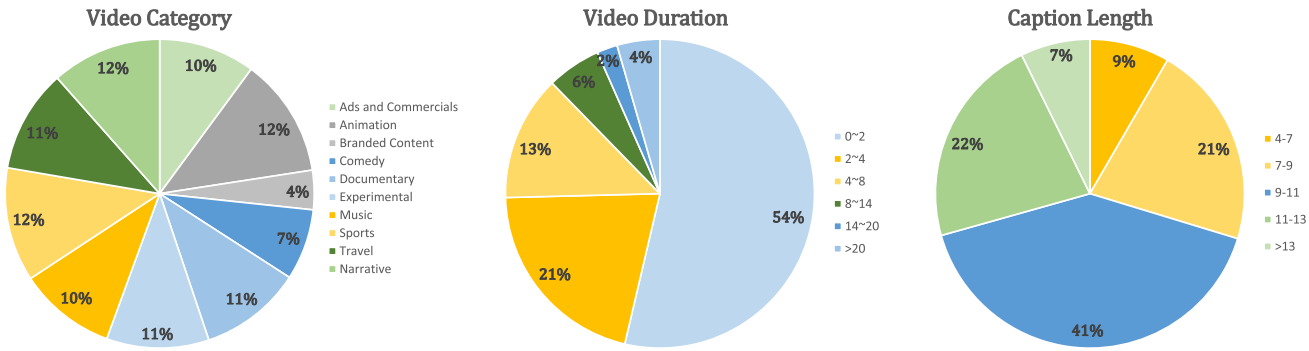
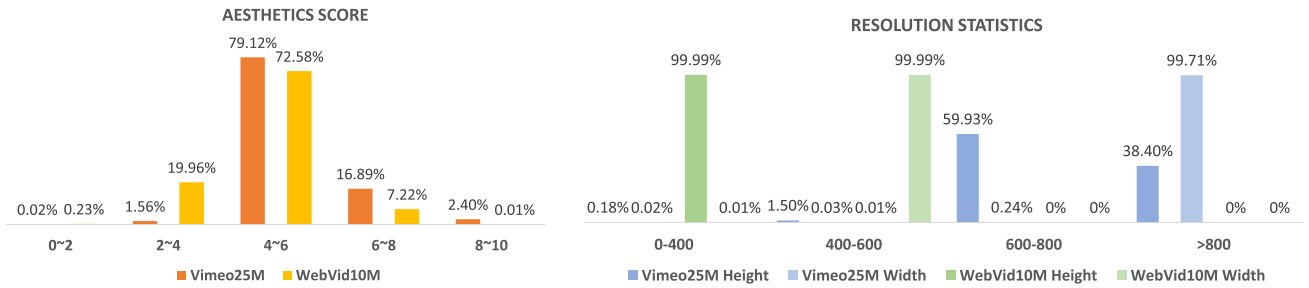


Fig. 7 Vimeo25M general information statistics. We show statistics of video categories, clip durations, and caption word lengths in Vimeo25M



(a) Aesthetics score statistics

(b) Resolution statistics

Fig. 8 Aesthetics score, video resolution statistics. We compare Vimeo25M with WebVid10M in terms of **a** aesthetics score and **b** video spatial resolution

learning (Bengio et al., 2009). In the initial stage, we utilize WebVid10M as the primary video data source, along with Laion5B, as the content within these videos is relatively simpler compared to the other dataset. Subsequently, we gradually introduce Vimeo25M to train the model on more complex scenes, subjects, and motion. In addition, we filter out an internal high-quality video-text dataset which contains 2000 videos for quality fine-tuning.

Temporal Interpolation model is initialized from our pre-trained base T2V model. In order to accommodate our concatenated inputs of high and low frame-rate frames, we extend the architecture by incorporating an additional convolutional layer. During training, we utilize WebVid10M as the primary dataset. In the later stages of training, we gradually introduce Vimeo25M, which allows us to leverage its watermark-free videos, thus assisting in eliminating watermarks in the interpolated output. While patches of dimensions 256×256 are utilized during training, the trained model can successfully interpolate base videos at a resolution of 320×512 during inference.

The spatial layers of our VSR model is initialized from the pre-trained diffusion-based image $\times 4$ upscaler, keeping these layers fixed throughout training. Only the newly inserted temporal layers, including temporal attention and 3D CNN layers, are trained. Similar to the base model, we employ the WebVid10M and Laion5B (with resolution ≥ 1024) datasets for joint image-video training. To facilitate this, we transform the image data into video clips by applying random translations to simulate handheld camera movements. For training purposes, all videos and images are cropped into patches of size 320×320. Once trained, the model can effectively process videos of arbitrary sizes, offering enhanced results.

At inference stage, our base model requires 10 s to produce a 16-frame video, interpolation model requires 180 s to generate 61 frames, and VSR model requires 6 min for video enhancement. In terms of memory consumption, our base, interpolation and VSR models require 13GB, 40GB and 42GB respectively.

5.3 Qualitative Evaluation

We present qualitative results of our approach through diverse text descriptions illustrated in Fig. 1. LaVie demonstrates its capability to synthesize videos with a wide range of content, including animals, movie characters, and various objects. Notably, our model exhibits a strong ability to combine spatial and temporal concepts, as exemplified by the synthesis of actions like *"Yoda playing guitar,"*. These results indicate that our model learns to compose different concepts by capturing the underlying distribution rather than simply memorizing the training data.

Furthermore, we compare our generated results with three state-of-the-art and showcases the visual quality comparison in Fig. 9. LaVie outperforms Make-A-Video in terms of visual fidelity. Regarding the synthesis in the *"Van Gogh style"*, we observe that LaVie captures the style more effectively than the other two approaches. We attribute this to two factors: 1) initialization from a pretrained LDM facilitates the learning of spatio-temporal joint distribution, and 2) joint image-video fine-tuning mitigates catastrophic forgetting observed in Video LDM and enables knowledge transfer from images to videos more effectively. However, due to the unavailability of the testing code for the other two approaches, conducting a systematic and fair comparison is challenging.

5.4 Quantitative Evaluation

We perform a zero-shot quantitative evaluation on two benchmark datasets, UCF101 (Soomro et al., 2012) and MSR-VTT (Chen et al., 2021), to compare our approach with existing methods. However, due to the time-consuming nature of sampling a large number of high-definition videos (e.g., $\sim$10000) using diffusion models, we limit our evaluation to using videos from the base models to reduce computational duration. Additionally, we observed that current evaluation metrics FVD may not fully capture the real quality of the generated videos. Therefore, to provide a comprehensive assessment, we conduct a large-scale human evaluation to compare the performance of our approach with state-of-the-art. In addition, we leverage a novel video generation evaluation benchmark suit VBench (Huang et al., 2023) to thoroughly analyze the generated video quality of LaVie and state-of-the-art.

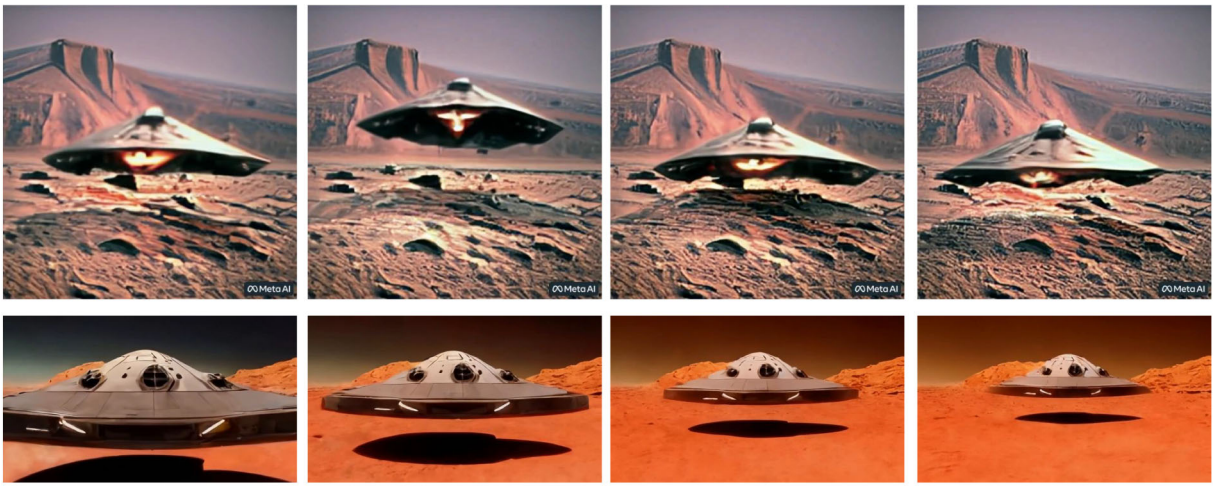
Zero-shot Evaluation on UCF101. We evaluate the quality of the synthesized results on UCF-101 dataset using the FVD, following the approach of TATS by employing the pretrained I3D (Carreira & Zisserman, 2017) model as the backbone. Similar to the methodology proposed in Video LDM, we utilize class names as text prompts and generate 100 samples per class, resulting in a total of 10,100 videos. During video sampling and evaluation, we generate 16 frames per video with a resolution of 320×512. Each frame is then center-cropped to a square size of 270×270 and resized to 224×224 to fit the I3D model input requirements.

The results, presented in Table 1, demonstrate that our model outperforms all baseline methods, except for Make-A-Video. However, it is important to note that we utilize a smaller training dataset (WebVid10M+Vimeo25M) compared to Make-A-Video, which employs WebVid10M and HD-VILA-100 M for training. Furthermore, in contrast to Make-A-Video, which manually designs a template sentence for each class, we directly use the class name as the text prompt, following the approach of Video LDM. When considering methods with the same experimental setting, our approach outperforms the state-of-the-art result of Video LDM by 24.31, highlighting the superiority of our method and underscoring the importance of the proposed dataset for zero-shot video generation (Table 2).

Zero-shot Evaluation on MSR-VTT. For the MSR-VTT dataset, we conduct our evaluation by randomly selecting one caption per video from the official test set, resulting in a total of 2,990 videos. We assess the text-video semantic similarity using the clip similarity (CLIPSIM) metric. To compute CLIPSIM, we calculate the clip text-image similarity for each frame, considering the given text prompts, and then calculate the average score. In this evaluation, we employ the ViT-B-32 clip model as the backbone, following the methodology outlined in previous work (Blattmann et al., 2023) to ensure a fair comparison. Our experimental setup and details are consistent with the previous work. As shown in Table 2, the results demonstrate that LaVie achieves superior or competitive performance compared to state-of-the-art methods, highlighting the effectiveness of our proposed training scheme and the utilization of the Vimeo25M dataset. These findings underscore

the efficacy of our approach in capturing text-video semantic similarity.

Human Evaluation. Deviating from previous methods that primarily focus on evaluating general video quality, we contend that a more nuanced assessment is necessary to comprehensively evaluate the generated videos from various perspectives. In light of this, we compare our method with two existing approaches, VideoCrafter and ModelScope, leveraging the accessibility of their testing platforms. To conduct a thorough evaluation, we enlist the assistance of 30



(a) Make-A-Video (top) & ours (bottom). *"Hyper-realistic spaceship landing on mars."*

(b) VideoLDM (top) & ours (bottom). *"A car moving on an empty street, rainy evening, Van Gogh painting."*

(c) Imagen Video (top) & ours (bottom). *"A cat eating food out of a bowl in style of Van Gogh."*

Fig. 9 Comparison with state-of-the-art methods. We compared to **a** Make-A-Video, **b** Video LDM and **c** Imagen Video. In each sub-figure, *bottom row* shows our result. We compare with Make-A-Video at spatial-resolution 512 × 512 and with the other two methods at 320 × 512

Table 1 Comparison with SoTA *w.r.t.* FVD for zero-shot T2V generation on UCF101

Methods	Pretrain on image	Image generator	Resolution	FVD (↓)
CogVideo (Chinese) (Hong et al., 2023)	No	CogView	480 × 480	751.34
CogVideo (English) (Hong et al., 2023)	No	CogView	480 × 480	701.59
Make-A-Video (Singer et al., 2023)	No	DALL·E2	256 × 256	367.23
VideoFusion (Luo et al., 2023)	Yes	DALL·E2	256 × 256	639.90
Magic Video (Zhou et al., 2022)	Yes	Stable Diffusion	256 × 256	699.00
LVDM (He et al., 2022)	Yes	Stable Diffusion	256 × 256	641.80
Video LDM (Blattmann et al., 2023)	Yes	Stable Diffusion	320 × 512	550.61
PYoCo (Ge et al., 2023)	Yes	eDiff-I	512 × 512	355.19
Ours (w/o Vimeo25M)	Yes	Stable Diffusion	320 × 512	540.30
Ours	Yes	Stable Diffusion	320 × 512	**350.00**

Table 2 Comparison with SoTA *w.r.t.* CLIPSIM for zero-shot T2V generation on MSR-VTT

Methods	Zero-Shot	CLIPSIM (↑)
GODIVA (Wu et al., 2021)	No	0.2402
NÜWA (Wu et al., 2022)	No	0.2439
CogVideo (Chinese) (Hong et al., 2023)	Yes	0.2614
CogVideo (English) (Hong et al., 2023)	Yes	0.2631
Make-A-Video (Singer et al., 2023)	Yes	0.3049
VideoLDM (Blattmann et al., 2023)	Yes	0.2929
ModelScope (Wang et al., 2023)	Yes	0.2930
Ours	Yes	0.2949

Table 3 Human Preference on overall video quality

Metrics	Ours > ModelScope	Ours > VideoCrafter	ModelScope > VideoCrafter
Video quality	75.00%	75.58%	59.10%

Table 4 Human Evaluation on five pre-defined metrics. Each number signifies the proportion of examiners who voted for a particular category (good, normal, or bad) out of all votes

	VideoCrafter			ModelScope			Ours		
Metrics	Bad	Normal	Good	Bad	Normal	Good	Bad	Normal	Good
Motion Smoothness	0.24	0.58	0.18	0.16	0.53	0.31	0.20	0.45	**0.35**
Motion Reasonableness	0.53	0.33	0.14	0.37	0.40	0.22	0.40	0.32	**0.27**
Subject Consistency	0.25	0.40	0.35	0.18	0.34	0.48	0.15	0.26	**0.58**
Background Consistency	0.10	0.40	0.50	0.08	0.28	0.63	0.06	0.22	**0.72**
Face/Body/Hand quality	0.69	0.24	0.06	0.51	0.31	0.18	0.46	0.30	**0.24**

Bold values indicate the best performance

human raters and employ two types of assessments. Firstly, we ask the raters to compare pairs of videos in three different scenarios: ours *v.s.* ModelScope, ours *v.s.* VideoCrafter, and ModelScope *v.s.* VideoCrafter. Raters are instructed to evaluate the overall video quality to vote which video in the pair has better quality. Secondly, we request raters to evaluate each video individually using five pre-defined metrics: motion smoothness, motion reasonableness, subject consistency, background consistency, and face, body, and hand quality. Raters are required to assign one of three labels, "good", "normal", or "bad" for each metric. All human studies are conducted without time limitations.

As presented in Tables 3 and 4, our proposed method surpasses the other two approaches, achieving the highest preference among human raters. However, it is worth noting that all three approaches struggle to achieve a satisfactory score in terms of "motion smoothness" indicating the ongoing challenge of generating coherent and realistic motion. Furthermore, producing high-quality face, body, and hand visuals remains challenging.

Table 5 VBench Evaluation. We show comparison across each of the 16 VBench dimensions. Higher scores indicate relatively better performance

Models	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
ModelScope	89.87%	95.29%	98.28%	95.79%	66.39%	52.06%	58.57%	82.25%
VideoCrafter	86.24%	92.88%	97.60%	91.79%	**89.72%**	44.41%	57.22%	87.34%
CogVideo	92.19%	95.42%	97.64%	96.47%	42.22%	38.18%	41.03%	73.40%
AnimateDiff	95.3%	97.68%	**98.75%**	**97.76%**	40.83%	**67.16%**	**70.1%**	90.9%
VideoCrafter2	**96.85%**	**98.22%**	98.41%	97.73%	42.5%	63.13%	67.22%	92.55%
Latte	91.61%	95.98%	97.94%	96.45%	68.33%	63.35%	66.69%	88.01%
LaVie (pretrained)	91.41%	97.47%	98.30%	96.38%	49.72%	54.94%	61.90%	91.82%
LaVie (QT)	94.13%	96.32%	97.81%	96.12%	60.00%	62.18%	66.04%	**94.18%**

Models	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
ModelScope	38.98%	92.40%	81.72%	33.68%	39.26%	23.39%	25.37%	25.67%
VideoCrafter	25.93%	93.00%	78.84%	36.74%	43.36%	21.57%	25.42%	25.21%
CogVideo	18.11%	78.20%	79.57%	18.24%	28.24%	22.01%	7.80%	7.70%
AnimateDiff	36.88%	92.6%	87.47%	34.6%	50.19%	22.42%	**26.03%**	27.04%
VideoCrafter2	40.66%	95.0%	**92.92%**	35.86%	**55.29%**	**25.13%**	25.84%	**28.23%**
Latte	35.44%	90.0%	85.73%	41.43%	36.18%	24.03%	24.80%	26.80%
LaVie (pretrained)	33.32%	96.80%	86.39%	34.09%	52.69%	23.56%	25.93%	26.41%
LaVie (QT)	**44.66%**	**97.40%**	87.65%	**45.63%**	54.68%	24.33%	25.12%	26.98%

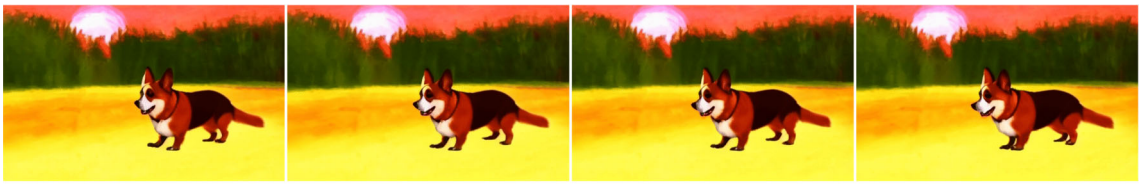
Bold values indicate the best performance

VBench Evaluation. We apply VBench on pre-trained and quality fine-tuned (QT) models to analyze the performance of using two-stage training strategies. Table 5 shows that after pre-training, LaVie has already outperformed state-of-the-art in many dimensions. After quality fine-tuning, the visual quality of generated videos has been further boosted, in particular "subject consistency", "aesthetic quality" and "imaging quality", which can also been observed in Fig. 10. It proves that fine-tuned the pre-trained T2V model on small-scale high-quality video dataset is an effective way to improve the generated quality.

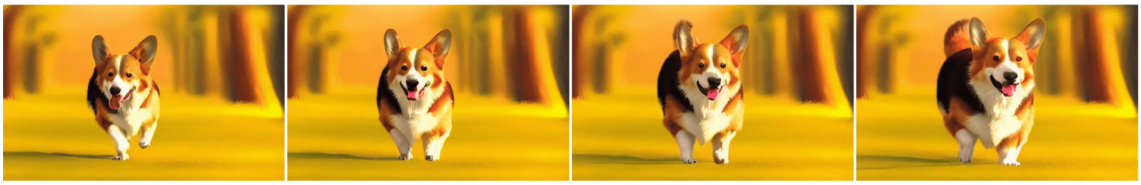
5.5 Further Analysis

Training scheme analysis. We conduct a qualitative analysis of the training scheme employed in our experiments, as well as the performance by using different temporal modules. We compare our joint image-video fine-tuning approach with two other experimental settings: 1) fine-tuning the entire UNet architecture based on WebVid10M, and 2) training temporal modules while keeping the rest of the network frozen. The results, depicted in Fig. 11, highlight the advantages of our proposed approach. When fine-tuning the entire model on video data, we observe catastrophic forgetting. The concept of *"teddy bear"* gradually diminishes and the quality of its representation deteriorates significantly. Since the training videos contain very few instances of *"teddy bear"*, the model gradually adapts to the new data distribution, resulting in a loss of prior knowledge. In the second setting, we encounter difficulties in aligning the spatial knowledge from the image dataset with the newly learned temporal information from the video dataset. The significant distribution gap between the image and video datasets poses a challenge in effectively integrating the spatial and temporal aspects. The attempts made by the high-level temporal modules to modify the spatial distribution adversely affect the quality of the generated videos. In contrast, our joint image-video fine-tuning scheme effectively learns the joint distribution of image and video data. This enables the model to recall knowledge from the image dataset and apply the learned motion from the video dataset, resulting in higher-quality synthesized videos. The ability to leverage both datasets enhances the overall performance and quality of the generated results.

Temporal module analysis. We have conducted small-scale experiments on UCF101 for comparison. We explore three different settings. 1) Our proposed temporal module, 2) replacing RoPE in our module with absolute PE, 3) spatial-temporal self-attention used in Singer et al. (2023). We train three models with the same iterations and report FVD in Table 6. While 3) is much complex than our proposed temporal module, we didn't find improvement in visual quality and training is relatively slower. Small-scale quantitative evaluation demonstrates the effectiveness of our proposed temporal module.



A corgi walking in the park at sunset, oild painting style. (LaVie-pretrained)



A corgi walking in the park at sunset, oild painting style. (LaVie-QT))

A cute raccoon playing guitar in a park, oild painting style. (LaVie-pretrained)

A cute raccoon playing guitar in a park, oild painting style. (LaVie-QT)

Fig. 10 Comparison between LaVie-pretrained and LaVie-QT. After high-quality fine-tuning, LaVie is able to generate videos with higher aesthetic level

(a) Training entire model (b) Training temporal modules (c) Joint image-video fine-tuning

Fig. 11 Training scheme comparison. We show image results based on **a** training the entire model, **b** training temporal modules, and **c** joint image-video fine-tuning, respectively

Table 6 Ablation study on temporal modules

Methods	LaVie	Absolute PE	ST Self-attention
FVD	611	625	635

RoPE analysis. We further analyze the effectiveness of using RoPE for extrapolation, and compare our proposed method with model using absolute PE. We train both models on 16 frames, and generate 32 frames for comparison. Results are shown in Fig 12, from which we can observe that model using RoPE performs much better than the one using absolute PE. The latter started to crash when the length of generated video longer than the training data. In addition, we quantitatively analyze 32-frame extrapolation using VBench. We report results on two dimensions, *subject consistency* and *motion smoothness*, in Table 7.

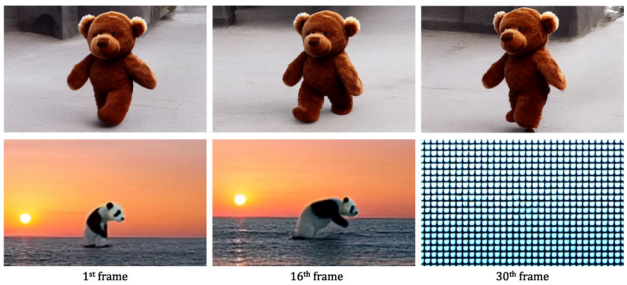


Fig. 12 Positional Embedding Comparison. We compare models trained on 16 frames using RoPE (up) and absolute positional embedding (bottom), and show extrapolation results. Model using absolute embedding starts to crash after the training length while model using RoPE performs well for extrapolation

Table 7 Quantitative analysis on extrapolation

Methods	Subject consistency	Motion smoothness
Absolute PE	0.90	0.62
RoPE	**0.96**	**0.93**

Bold values indicate the best performance

Comparison with Latte. As DiT-based model has proved to be effective for text-to-image/video generation, we also qualitatively and quantitatively compare our method with a transformer-based T2V model Latte, and show results in Fig. 13 and Table 5 respectively in the revised manuscript. We observed that Latte achieved better performance than LaVie on several temporal dimensions such as *temporal flickering*, *dynamic degree*, and *overall consistency*. We analyzed the architecture of two models and found that Latte contains more temporal modules (28 temporal transformer blocks) than LaVie, hence it has stronger capacity to model temporal signal. On spatial dimensions, Latte slightly outperforms LaVie on *aesthetic quality* and *imaging quality*. We conclude such differences might due to the differences in training data quality and pretrained T2I model.

5.6 More Applications

In this section, we present two applications to showcase the capabilities of our pretrained models in downstream tasks: 1) long video generation, and 2) personalized T2V generation using LaVie.

Long video generation. To extend the video generation beyond a single sequence, we propose a simple recursive method. Similar to temporal interpolation network, we incorporate the first frame of a video into the input layer of a UNet. By fine-tuning the base model accordingly, we enable the utilization of the last frame of the generated video as a conditioning input during inference. This recursive approach allows us to generate an extended video sequence. Figure 14

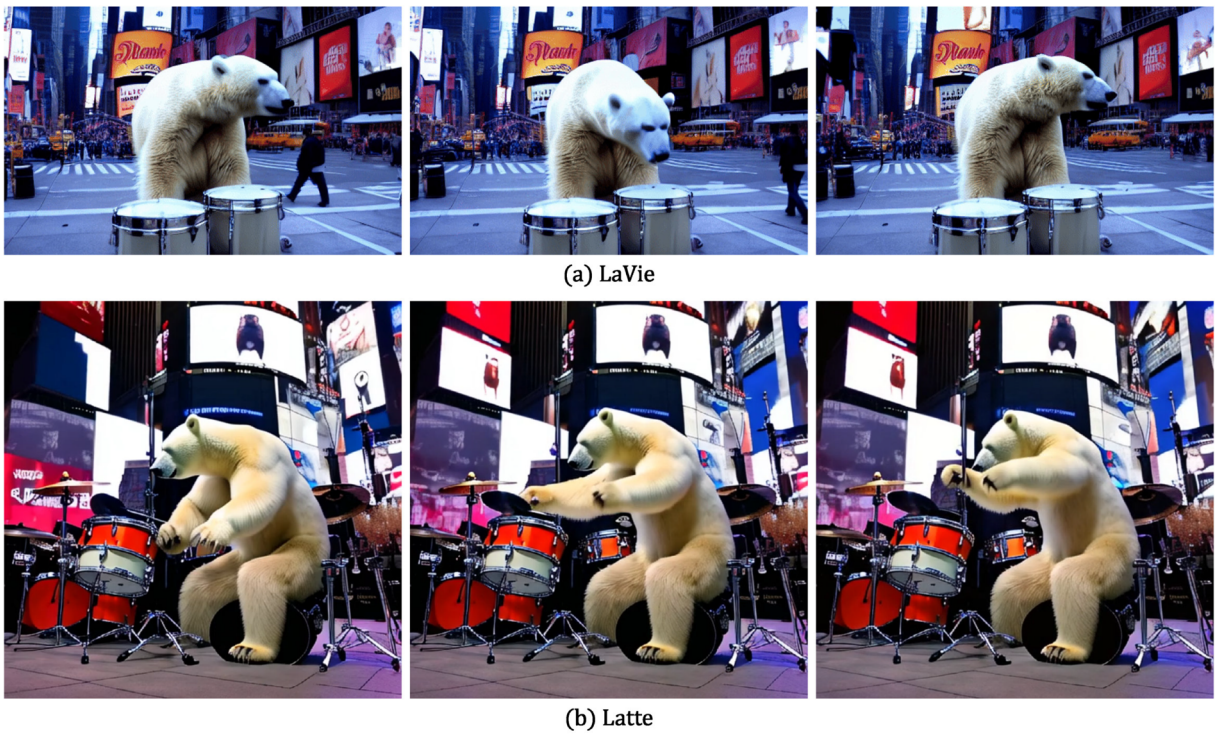
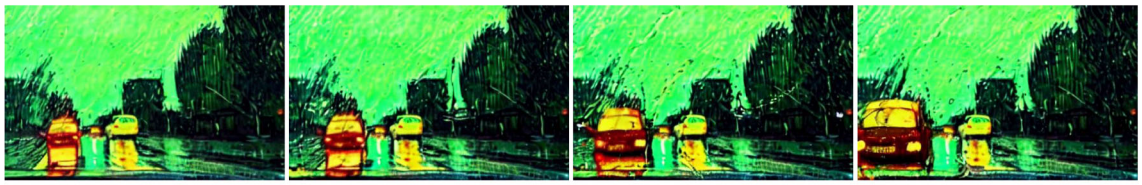


Fig. 13 Comparison with Latte. We show comparison with transformer-based video diffusion model Latte using prompt *'a polar bear playing drum kit in NYC Times Square, 4k, high resolution'*

A car moving on an empty street, rainy evening, Van Gogh painting. [0∼2s]

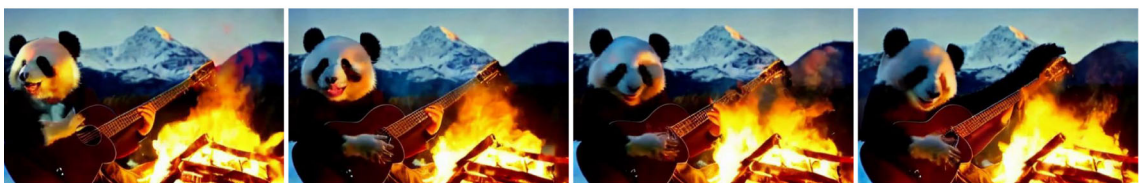
A car moving on an empty street, rainy evening, Van Gogh painting. [2∼4s]



A car moving on an empty street, rainy evening, Van Gogh painting. [4∼6s]

A panda playing guitar near a campfire, snow mountain in the background. [0∼2s]

A panda playing guitar near a campfire, snow mountain in the background. [2∼4s]



A panda playing guitar near a campfire, snow mountain in the background. [4∼6s]

Fig. 14 Long video generation. By employing autoregressive generation three times consecutively, we successfully extend the video length of our base model from 2 s to 6 s

(a) *Misaka Mikoto*

(b) *Misaka Mikoto walking in the city*

(c) *Misaka Mikoto in the space*

Fig. 15 Personalized T2V generation. We show results by adopting a LoRA-based approach in our model for personalized video generation. Samples used to train our LoRA are shown in (**a**). We use "Misaka Mikoto" as text prompts. Results from our video LoRA are shown in (**b**) and (**c**). By inserting pre-trained temporal modules into LoRA, we are able to animate "Misaka Mikoto" and control the results by combining them with different prompts

showcases the results of generating tens of video frames (excluding frame interpolation) using this recursive manner, applied five times. The results demonstrate that the quality of the generated video remains high, with minimal degradation in video quality. This reaffirms the effectiveness of our base model in generating visually appealing frames.

Personalized T2V generation. Although our approach is primarily designed for general text-to-video generation, we demonstrate its versatility by adapting it to personalized video generation through the integration of a personalized image generation approach, such as LoRA (Hu et al., 2022). In this adaptation, we fine-tune the spatial layers of our model using LoRA on self-collected images, while keeping the temporal modules frozen. As depicted in Fig. 15, the personalized video model for *"Misaka Mikoto"* is created after the fine-tuning process. The model is capable of synthesizing personalized videos based on various prompts. For instance, by providing the prompt *"Misaka Mikoto walking in the city"*, the model successfully generates scenes where *"Misaka Mikoto"* is depicted in novel places.

6 Limitations

While LaVie has demonstrated impressive results in general text-to-video generation, we acknowledge the presence of certain limitations. In this section, we highlight two specific challenges which are shown in Fig. 16:

Multi-subject generation: Our models encounter difficulties when generating scenes involving more than two subjects, such as *"Albert Einstein discussing an academic paper with Spiderman"*. There are instances where the model tends to mix the appearances of Albert Einstein and Spiderman, instead of generating distinct individuals. We have observed that this issue is also prevalent in the T2I model (Rombach et al., 2022). One potential solution for

(a) *Albert Einstein discussing an academic paper with Spiderman.*

(b) *Albert Einstein playing the violin.*

Fig. 16 Limitations. We show limitations on **a** mutiple-object generation and **b** failure of hands generation

improvement involves replacing the current language model, CLIP (Radford et al., 2021), with a more robust language understanding model like T5 (Raffel et al., 2020). This substitution could enhance the model's ability to accurately comprehend and represent complex language descriptions, thereby mitigating the mixing of subjects in multi-subject scenarios.

Hands generation: Generating human bodies with high-quality hands remains a challenging task. The model often struggles to accurately depict the correct number of fingers, leading to less realistic hand representations. A potential solution to address this issue involves training the model on a larger and more diverse dataset containing videos with human subjects. By exposing the model to a wider range of hand appearances and variations, it could learn to generate more realistic and anatomically correct hands.

7 Conclusion

In this paper, we present **LaVie**, a text-to-video foundation model that produces high-quality and temporally coherent results. Our approach leverages a cascade of video diffusion models, extending a pre-trained LDM with simple designed temporal modules enhanced by Rotary Position Encoding (RoPE). To facilitate the generation of high-quality and diverse content, we introduce Vimeo25M, a novel and extensive video-text dataset that offers higher resolutions and improved aesthetics scores. By jointly fine-tuning on both image and video datasets, LaVie demonstrates a remarkable capacity to compose various concepts, including styles, characters, and scenes. We conduct comprehensive quantitative and qualitative evaluations for zero-shot text-to-video generation, which convincingly validate the superiority of our method over state-of-the-art approaches. Furthermore, we showcase the versatility of our pre-trained base model in two additional tasks i.e. long video generation and personalized video generation. These tasks serve as additional evidence of the effectiveness and flexibility of LaVie. We envision LaVie as an initial step towards achieving high-quality T2V generation. Future research directions involve expanding the capabilities of LaVie to synthesize longer videos with intricate transitions and movie-level quality, based on script descriptions.

Acknowledgements This work is partially supported by the National Key R&D Program of China under Grant No. 2022ZD0160102, National Natural Science Foundation of China under Grant No. 62102150, and the Science and Technology Commission of Shanghai Municipality under Grant No. 23QD1400800 and No. 23YF1461900.

Data Availability The datasets used during and analyzed during the current study are available in the following public domain resources: **WebVid10M** (Bain et al., 2021) https://github.com/m-bain/webvid, **UCF101** (Soomro et al., 2012) https://www.crcv.ucf.edu/data/UCF101.php. The models and source data generated during and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethical Approval We acknowledge the ethical concerns that are shared with other T2I and T2V diffusion models. We aim to synthesize high-quality videos by giving text descriptions. Our approach can be used for movie production, making video games, artistic creation, generating synthetic data for other computer vision tasks, etc. We note that our framework has the potential to introduce unintended bias as a result of the training data.

References

Bain, M., Nagrani, A., Varol, G.. & Zisserman, A. (2021). Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al. (2022). ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*

Bhagat, S., Uppal, S., Yin, Z. & Lim, N. (2020). Disentangling multiple features in video sequences using gaussian processes in variational autoencoders. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., & Letts, A., et al. (2023). 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., & Kreis, K. (2023). Align your latents: High-resolution

video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Brock, A., Donahue, J. & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *ICLR*

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., & Luhman, E., et al. (2024). Video generation models as world simulators

Brooks, T., Hellsten, J., Aittala, M., Wang, T.-C., Aila, T., Lehtinen, J., Liu, M.-Y., Efros, A. A., & Karras, T. (2022). Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems, 35*, 31769–31781.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*

Chan, K.C.K., Zhou, S., Xu, X. & Loy, C.C. (2022). Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Chan, K.C.K., Zhou, S., Xu, X., & Loy, C.C. (2022). BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Chen, H., Li, J., Frintrop, S., & Hu, X. (2021). The msr-video to text dataset with clean annotations. arXiv preprint arXiv:2102.06448

Chen, J., YU, J., GE, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., & Li, Z. (2024). Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*

Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., & Shan, Y. (2024). Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7310–7320.

Chen, X., Xu, C., Yang, X., & Tao, D. (2020). Long-term video prediction via criticization and retrospection. *IEEE Transactions on Image Processing, 29*, 7090–7103.

Clark, A., Donahue, J., & Simonyan, K. (2019). Adversarial video generation on complex datasets. arXiv preprint arXiv:1907.06571

Dai, X., Hou, J., Ma, C.-Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., & Dubey, A., et al. (2023). Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F. & *et al.* (2024). Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*

Esser, P., Rombach, R. & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.-B., & Parikh, D. (2022). Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European conference on computer vision*

Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.-B., Liu, M.-Y., & Balaji, Y. (2023). Preserve your own correlation: A noise prior for video diffusion models. arXiv preprint arXiv:2305.10474

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*

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

Gupta, A., Yu, L., Sohn, K., Gu, X., Hahn, M., Fei-Fei, L., Essa, I., Jiang, L., & Lezama, J. (2023). Photorealistic video generation with diffusion models. arXiv preprint arXiv:2312.06662

He, Y., Yang, T., Zhang, Y., Shan, Y., & Chen, Q. (2022). Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221

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

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D.J. (2022). Video diffusion models. arXiv preprint arXiv:2204.03458

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *NeurIPS, 33*, 6840.

Hong, W., Ding, M., Zheng, W., Liu, X., & Tang, J. (2023). Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *ICLR*

Huang, Z., Chan, K.C.K., Jiang, Y., & Liu, Z. (2023). Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

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

Jiang, Y., Chan, K.C., Wang, X., Loy, C.C., & Liu, Z. (2021). Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Jiang, Y., Chan, K.C., Wang, X., Loy, C.C., & Liu, Z. (2022). Reference-based image and video super-resolution via c^2-matching. In *IEEE transactions on pattern analysis and machine intelligence*

Jiang, Y., Yang, S., Koh, T.L., Wu, W., Loy, C.C., & Liu, Z. (2023). Text2performer: Text-driven human video generation. arXiv preprint arXiv:2303.13495

Karras, T., Laine, S. & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Kingma, D.P., & Welling, M. (2014). Auto-encoding variational bayes. In *ICLR*

Li, Y., & Mandt, S. (2018). Disentangled sequential autoencoder. In *ICML*

Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., & Qiao, Y. (2023). VideoChat: Chat-Centric Video Understanding

Lu, H., Yang, G., Fei, N., Huo, Y., Lu, Z., Luo, P. & Ding, M. (2023). Vdt: General-purpose video diffusion transformers via mask modeling. In *ICLR*

Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., & Tan, T.-P. (2023). Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*

Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.-F., Chen, C., & Qiao, Y. (2024). Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048

Mokady, R., Hertz, A., Aberman, K., Pritch, Y., & Cohen-Or, D. (2022). Null-text inversion for editing real images using guided diffusion models. arXiv preprint arXiv:2211.09794

Nichol, A.Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In: *International Conference on Machine Learning*

Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., & Zhu, J.-Y. (2023). Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 conference proceedings*

Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*

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 *International conference on machine learning*

Radford, A., Metz, L. & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & 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., & Sutskever, I. (2021). Zero-shot text-to-image generation. In *International conference on machine learning*

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

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. *Advances in Neural Information Processing Systems, 35*, 36479–36494.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., & Norouzi, M. (2022). Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*, 4713.

Saito, M., Matsumoto, E., & Saito, S. (2017). Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems, 35*, 25278–25294.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., & Taigman, Y. (2023). Make-a-video: Text-to-video generation without text-video data. In *ICLR*

Skorokhodov, I., Tulyakov, S., & Elhoseiny, M. (2022). Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. In *ICLR*

Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *ICLR*

Soomro, K., Zamir, A.R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2021). Roformer: Enhanced transformer with rotary position embedding. arXiv preprint arXiv:2104.09864

Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D.N., & Tulyakov, S. (2021). A good image generator is what you need for high-resolution video synthesis. In *ICLR*

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., & Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971

Tulyakov, S., Liu, M.-Y., Yang, X., & Kautz, J. (2018). MoCoGAN: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*

Van Den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. Advances in neural information processing systems.

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating videos with scene dynamics. In *NeurIPS*

Wang, Y. (2021). *Learning to Generate Human Videos*. Theses: Inria—Sophia Antipolis; Université Cote d'Azur

Wang, Y., Bilinski, P., Bremond, F. & Dantcheva, A. (2020). Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*

Wang, Y., Bilinski, P., Bremond, F., & Dantcheva, A. (2020). G3AN: Disentangling appearance and motion for video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Wang, Y., Bremond, F., & Dantcheva, A. (2021). Inmodegan: Interpretable motion decomposition generative adversarial network for video generation. arXiv preprint arXiv:2101.03049

Wang, Y., Ma, X., Chen, X., Dantcheva, A., Dai, B., & Qiao, Y. (2023). Leo: Generative latent image animator for human video synthesis. arXiv preprint arXiv:2305.03989

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., & Zhang, S. (2023). Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571

Wang, J., Yue, Z., Zhou, S., Chan, K.C., & Loy, C.C. (2023). Exploiting diffusion prior for real-world image super-resolution. arXiv preprint arXiv:2305.07015

Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., & Duan, N. (2021). Godiva: Generating open-domain videos from natural descriptions. arXiv preprint arXiv:2104.14806

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., & Duan, N. (2022). Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*

Xie, J., Gao, R., Zheng, Z., Zhu, S.-C. & Wu, Y.N. (2020). Motion-based generator model: Unsupervised disentanglement of appearance, trackable and intrackable motions in dynamic patterns. In *Proceedings of the AAAI conference on artificial intelligence*

Yan, W., Zhang, Y., Abbeel, P., & Srinivas, A. (2021). Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., & Feng, G., et al. (2024). Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072

Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.-W., & Shin, J. (2022). Generating videos with dynamics-aware implicit generative adversarial networks. In *ICLR*

Zhang, L., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543

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

Zhang, Y., Xing, Z., Zeng, Y., Fang, Y., & Chen, K. (2024). Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Zhang, Q., Yang, C., Shen, Y., Xu, Y., & Zhou, B. (2023). Towards smooth video composition. In *ICLR*

Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., & Feng, J. (2022). Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018

Zhou, S., Chan, K., Li, C., & Loy, C. C. (2022). Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems, 35*(2022), 30599–30611.

Zhou, S., Zhang, J., Zuo, W., & Loy, C. C. (2020). Cross-scale internal graph neural network for image super-resolution. *Advances in Neural Information Processing Systems, 33*, 3499–3509.