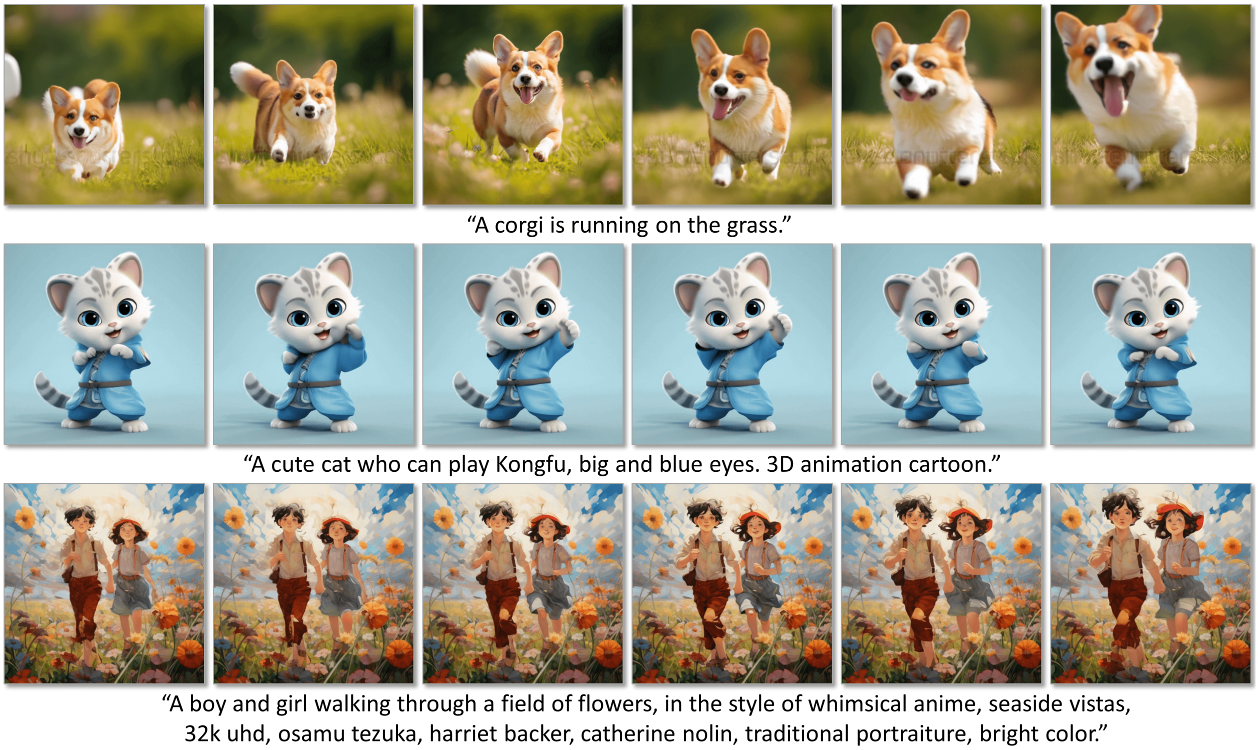
MicroCinema: A Divide-and-Conquer Approach for Text-to-Video Generation

Yanhui Wang[1,2*†], Jianmin Bao[2*], Wenming Weng[1], Ruoyu Feng[1], Dacheng Yin[1], Tao Yang[3],
Jingxu Zhang[1], Qi Dai[2], Zhiyuan Zhao[2], Chunyu Wang[2], Kai Qiu[2], Yuhui Yuan[2],
Xiaoyan Sun[1], Chong Luo[1,2‡], Baining Guo[2]
[1]University of Science and Technology of China [2]Microsoft Research Asia [3]Xi'an Jiaotong University
[*]Equal Contribution [†]This work was done during the internship at MSRA [‡]Corresponding author.


"A corgi is running on the grass."

"A cute cat who can play Kongfu, big and blue eyes. 3D animation cartoon."

"A boy and girl walking through a field of flowers, in the style of whimsical anime, seaside vistas, 32k uhd, osamu tezuka, harriet backer, catherine nolin, traditional portraiture, bright color."

Figure 1. Sample videos produced by MicroCinema, our proposed text-to-video generation system. They showcase MicroCinema's ability to create coherent and high-quality videos, with precise motion aligned with text prompts. Image reference generated by Midjourney.

Abstract

*We present MicroCinema, a straightforward yet effective framework for high-quality and coherent text-to-video generation. Unlike existing approaches that align text prompts with video directly, MicroCinema introduces a Divide-and-Conquer strategy which divides the text-to-video into a two-stage process: text-to-image generation and image&text-to-video generation. This strategy offers two significant advantages. a) It allows us to take full advantage of the recent advances in text-to-image models, such as Stable Diffusion, Midjourney, and DALLE, to generate photorealistic and highly detailed images. b) Leveraging the generated image, the model can allocate less focus to fine-grained appearance details, prioritizing the efficient learning of motion dynamics. To implement this strategy effectively, we introduce two core designs. First, we propose the Appearance Injection Network, enhancing the preservation of the appearance of the given image. Second, we introduce the Appearance Noise Prior, a novel mechanism aimed at maintaining the capabilities of pre-trained 2D diffusion models. These design elements empower MicroCinema to generate high-quality videos with precise motion, guided by the provided text prompts. Extensive experiments demonstrate the superiority of the proposed framework. Concretely, MicroCinema achieves **SOTA** zero-shot FVD of **342.86** on UCF-101 and **377.40** on MSR-VTT.*

1. Introduction

Diffusion models [13, 37] have achieved remarkable success in text-to-image generation, such as DALL-E [29], Stable Diffusion [32], Imagen [33], among others. They can generate unseen image content based on novel text concepts, showcasing impressive capabilities for image content generation and manipulation. Consequently, researchers have sought to extend the success of diffusion models to text-to-video generation.

One prevalent strategy involves training large-scale text-to-video diffusion models directly [14, 15, 36, 40]. These models employ cascade spatiotemporal diffusion models to learn from text and video pairs. While capable of producing high-quality videos, they pose challenges due to substantial GPU resource requirements and the need for extensive training data. Recently, some works [5, 7] have presented a cost-effective strategy. These methods entail the insertion of temporal layers into a text-to-image model, followed by fine-tuning on paired text and video data to create a text-to-video model. However, videos generated using this approach may encounter issues related to appearance and temporal coherence. We argue that maintaining appearance and temporal coherence is crucial for effective video generation.

In this paper, we present a novel approach, named MicroCinema, which employs a divide-and-conquer strategy to address appearance and temporal coherence challenges in video generation. The model features a two-stage generation pipeline. In the first stage, we generate a center frame, which serves as the foundation for subsequent video clip generation based on the input text. This design offers the flexibility to utilize any existing text-to-image generator for the initial stage, allowing users to incorporate their own images to establish the desired scene.

The second stage, known as image&text-to-video, concentrates on motion modeling. To achieve this, we leverage the open-source text-to-image generation model called Stable Diffusion (SD) [32] and inject temporal layers into it to obtain a three-dimensional (3D) network structure. The SD model has been trained on the filtered large-scale LAION dataset [35]. Its strong performance in generating high-quality images demonstrates its ability to capture spatial information within visual signals. To further enhance the model's ability to capture motion, we propose two core designs for the image&text-to-video model.

First, we introduce an Appearance Injection Network to inject the given image as a condition to guide the video generation. Concretely, it shares the structure of the encoder and middle part of the 3D U-Net and feeds the learned feature into the main branch via dense injection in a multi-layer manner. The dense injection operation better injects the appearance into the main branch, thus releasing the model from appearance modeling and encouraging the model dedicated to motion modeling. Second, we propose an appearance-aware noise strategy to preserve the pretrained capability of the SD model by modifying the i.i.d. noise in the diffusion process. Specifically, we add an appropriate amount of center frame to the i.i.d. noise without altering the overall diffusion training and inference process. This appearance-aware noise provides an intuitive cue to the model to generate a video whose appearance is similar to the given center frame, thereby unleashing its motion modeling capabilities.

Equipped with these designs, our framework can generate appearance-preserving and coherent videos with a given image and text. Extensive experiments demonstrate the superiority of MicroCinema. We achieve a **state-of-the-art** zero-shot FVD of **342.86** on UCF101 [38] and **377.40** on MSR-VTT [51] when training on the public WebVid-10M [4] dataset.

In summary, our contributions are presented as follows:

- We introduce an innovative two-stage text-to-video generation pipeline that capitalizes on a key-frame image generated by any off-the-shelf text-to-image generator in the initial stage. Subsequently, both the generated key-frame image and text serve as inputs for the video generation process in the second stage.
- We propose an Appearance Injection Network structure to encourage the 3D model to focus on motion modeling during the image&text-to-video generation process.
- We introduce an effective and distinctive Appearance Noise Prior tailored for fine-tuning text-to-image diffusion models. This modification significantly elevates the quality of video generation.
- In-depth quantitative and qualitative results are presented to validate the video generation capability of our proposed MicroCinema.

2. Related Work

The task of video generation involves addressing two fundamental challenges: image generation and motion modeling. Various approaches have been employed for image generation, including Generative Adversarial Networks (GANs) [2, 3, 8, 34], Variational autoencoder (VAE) [12, 31] and flow-based methods [6]. Recently, the state-of-the-art methods are built on top of diffusion models such as DALLE-2 [30], Stable Diffusion [32], GLIDE [24] and Imagen [33], which achieved impressive results. Extending these models for video generation is a natural progression, though it necessitates non-trivial modifications.

Text-to-Video Models. Image diffusion models adopt 2D U-Net with few exceptions [26]. To generate temporally smooth videos, temporal convolution (conv) or attention layers are also introduced. Notably, in Align-your-latents [5], 3D conv layers are interleaved with the existing spatial layers to align individual frames in a temporally consistent manner. This factorized space-time design has become the de facto standard and has been used in VDM [15], Imagen Video [14], and CogVideo [17]. Besides, it creates a concrete partition between the pre-trained two-dimensional (2D) conv layers and the newly initialized temporal conv layers, allowing us to train the temporal convolutions from scratch while retaining the previously learned knowledge in the spatial convolutions' weights. More recent work Latent-Shift [1] introduces no additional parameters but shifts channels of spatial feature maps along the temporal dimension, enabling the model to learn temporal coherence. Many approaches rely on temporal layers to implicitly learn motions from paired text and videos [5, 14, 15, 17, 36]. The generated motions, however, still lack satisfactory global coherence and fail to faithfully capture the essential movement patterns of the target subjects.

Leveraging Prior for Text-to-Video Diffusion Models. Generating natural motions poses a significant challenge in video generation. Many attempts are focused on leveraging prior into the text-to-video generation process. ControlVideo [54] directly utilizes ground truth motions, represented as depth maps or edge maps, as conditions for video diffusion models, demonstrating the importance of motion in video generation. GD-VDM [19] involves a two-phase generation process leveraging generating depth videos followed by a novel diffusion Vid2Vid model that generates a coherent real-world video in the autonomous driving scenario. However, it is not clear whether it can be applied to general scenes due to the lack of depth training data. Make-Your-Video [49] utilizes a standalone depth estimator to extract depth from a driving video, bypassing the need for depth generation, to generate new videos. In Leo [46], a motion diffusion model is trained to generate a sequence of motion latents, fed to a decoder network to recover the optical flows to animate the input image. Meanwhile, other methods involve linear displacement of codes in latent space [44], noise correlation [18], and generating textual descriptions for motion [16], serving as conditions for video generation models. More recent work PYoCo [7] proposes the video diffusion noise prior for a diffusion model and cost-effectively fine-tuning the text-to-image model.

Our proposed framework differs significantly from existing methods by employing a Divide-and-Conquer strategy. In our approach, we first generate images and subsequently capture motion dynamics along the temporal dimension. We also notice that a previous method Make-A-Video [36] has adopted a similar approach. However, our method introduces a novel model network design and incorporates an appearance-noise prior. This innovation ensures the generated video not only maintains the appearance established in the initial stage but also demonstrates superior motion modeling capabilities, a feature notably absent in Make-A-Video and concurrently related methods [20, 53].

3. MicroCinema

3.1. Overview

Our approach decomposes the text-to-video generation process into two distinct stages. Initially, we employ prevalent off-the-shelf text-to-image generation techniques to produce a key frame. Subsequently, both the key frame, acting as the center frame, and the text prompts are used as input to the image&text-to-video model to generate videos. We argue that the image&text-to-video model in a two-stage framework exhibits the potential for yielding more natural videos compared to the single-stage text-to-video model. This argument rests on the premise that by incorporating the center frame as a condition, our approach mitigates the model's burden in learning complicated appearance.

In the image&text-to-video generation stage, we adopt a cascaded approach to produce high-quality videos. First, we use a base image&text-to-video model to generate low frame rate videos from given image and text. Then, an adapted temporal interpolation model, derived from the base model, is employed to augment the frame rate. Finally, an off-the-shelf spatial super-resolution model is incorporated to render high-definition videos. This paper focuses on explaining the base model design and detailing its adaptation into the temporal interpolation model.

Base image&text-to-video model. Fig. 2 illustrates the overall architecture of the base image&text-to-video model in MicroCinema. This model is extended from the widely recognized Stable Diffusion (SD) model [32]. Following previous attempts [5, 36], we first extend the 2D U-Net into a 3D structure. We first enhance the original model by adding a 1D temporal convolution (conv) layer following each 2D spatial conv layer, enhancing its ability to handle temporal alignments. Additionally, we introduce a 1D temporal attention layer after every 2D spatial attention layer. These attention layers effectively capture long-range temporal correspondence, complementing the functionality of the 1D conv layers. To protect the strong capability of SD, we zero-initialize all the convolution and attention temporal layers and add a skip connection to it. Based on these modifications, we obtain a 3D model that can handle text-to-video generation. The base image&text-to-video model showcases two crucial innovations: the AppearNet and the appearance noise prior. Both are designed to incorporate appearance information from the key frame. A detailed explanation of these technical advancements will be provided in Sec. 3.2 and Sec. 3.3.

Temporal interpolation model. Our base model generates videos at a resolution of 320 × 320 pixels with a frame rate of 2 frames per second (fps). To enhance temporal quality, we train a temporal interpolation model designed for four-fold temporal super-resolution (TSR). This TSR model mirrors the architecture of the base model with slight modifica-

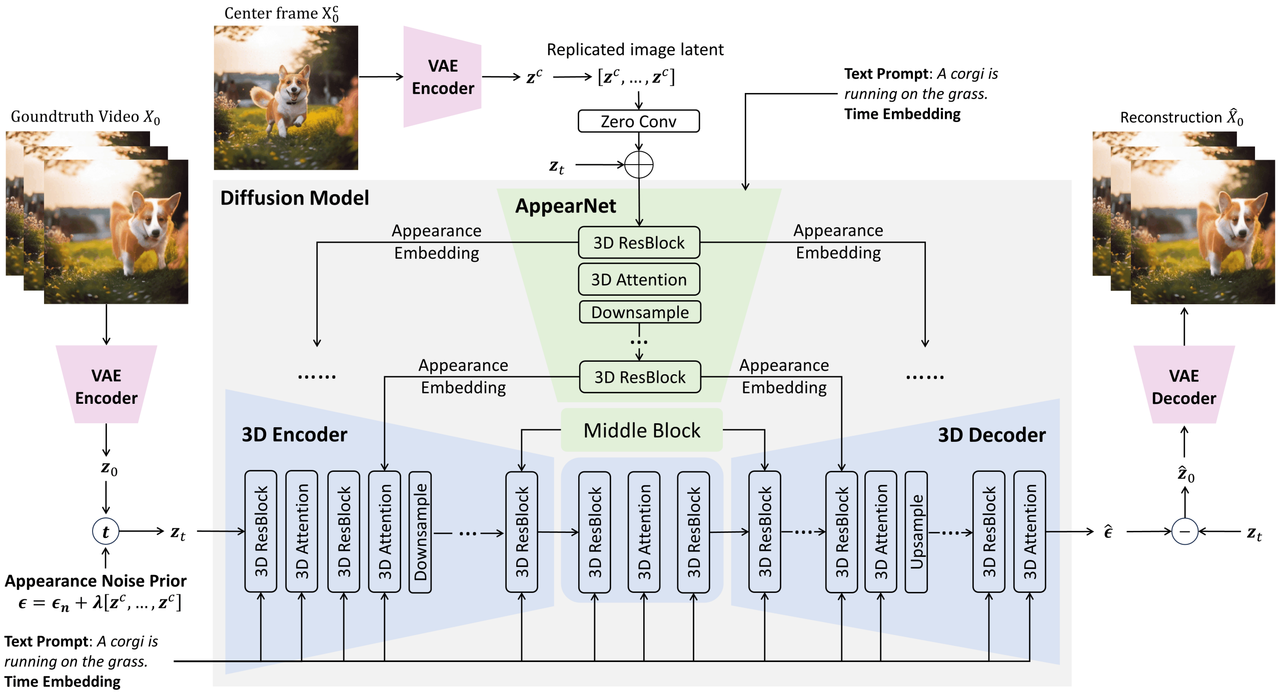


Figure 2. Overall architecture of our proposed diffusion-based image&text-to-video model in MicroCinema. The proposed AppearNet aims to provide appearance information for video generation.

tions. The base model employs only one conditional image (the center frame) while there are two conditional images (the start and end frames) in the TSR model. Accordingly, we alter the input of the AppearNet, shifting from duplicating the center frame to utilizing the interpolated latent representations of the given first and last frames. Leveraging this model consecutively on adjacent frames from previous steps boosts the frame rate from 2 fps to 32 fps.

3.2. Appearance Injection Network

To enhance the model's capability in handling reference center frame, we introduce the Appearance Injection Network, abbreviated as AppearNet, to the 3D network as depicted in Fig. 2. Inspired by ControlNet [5], we let AppearNet inherit the encoder and the middle part of the backbone network. Let N be the frame length of the output video. Then the center frame z^c is replicated for N times to create an image sequence, denoted as $[z^c, z^c, \ldots, z^c]$. It is used as input to the AppearNet to offer a robust appearance cue for generating output video frames.

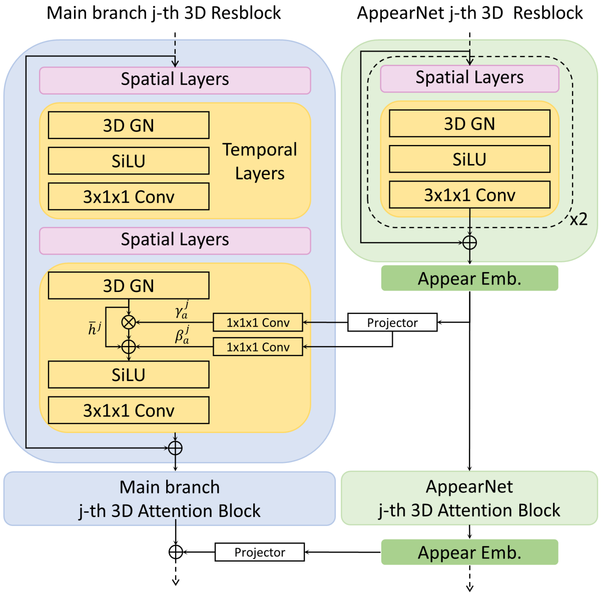


Figure 3. AppearNet injects multi-scale features into the main branch to perform a dense fusion.

We apply a multi-scale and dense fusion mechanism to seamlessly integrate the outputs of the AppearNet into the main branch. The multi-scale output of AppearNet is injected into both the encoder and the decoder of the main branch at the corresponding scales. In addition to the commonly used additive operation, we introduce an effective strategy of de-normalization [25] to inject the feature into the corresponding normalization layer of the main branch. As shown in Fig. 3, at the j-th feature level, let $\boldsymbol{h}_m^j$ denote the activation map in the main branch. Before integration, we perform 3D Group Normalization [48] on $\boldsymbol{h}_m^j$:

$$\bar{\boldsymbol{h}}^j = \frac{\boldsymbol{h}_m^j - \boldsymbol{\mu}^j}{\boldsymbol{\sigma}^j}. \tag{1}$$

Here $\boldsymbol{\mu}^j$ and $\boldsymbol{\sigma}^j$ are the means and standard deviations of $\boldsymbol{h}_m^j$'s group-wise activations. For AppearNet feature integration, let $\boldsymbol{f}_a^j$ be the AppearNet embedding on this feature level, we compute the output activation $\boldsymbol{o}^j$ by denormalizing the normalized $\bar{\boldsymbol{h}}^j$ according to $\boldsymbol{f}_a^j$, formulated as

$$\boldsymbol{o}^j = (\gamma_a^j + 1) \otimes \bar{\boldsymbol{h}}^j + \beta_a^j, \tag{2}$$

where γ_a^j and β_a^j are obtained by convolving from the feature map $\boldsymbol{f}_a^j$. The computed γ_a^j and β_a^j are multiplied and added to $\bar{\boldsymbol{h}}^j$ in an element-wise manner. Equipped with this design, our entire structure could better maintain the appearance from a given center frame while possessing the ability to generate videos based on text and image conditions.

3.3. Appearance Noise Prior

Fine-tuning from a text-to-image model proves to be a cost-effective approach for acquiring a video generation model. However, this process presents challenges due to the transition of the output space from images to videos. In the context of a typical T2I diffusion model, it tends to generate appearance-irrelevant images from a sequence of independent noise (sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$). In video generation, a sequence of independent noise should ideally yield a video with a coherent appearance. Therefore, the fine-tuning process may potentially compromise the capability of the original 2D T2I model. Our focus lies in preserving the effectiveness of the original 2D T2I model during the fine-tuning process for the image&text-to-video model.

For our proposed image&text-to-video model, the model should expand the given center image to a sequence of frames, which have a similar appearance to the center frame. Consequently, the output video is predominantly determined by the center frame rather than the sampled noise in the original diffusion process. To address this, we modify the noise distribution to align with the appearance of the given center frame. Leveraging the denoising property of the diffusion model, we introduce Appearance Noise Prior by adding an appropriate amount of the center frame into the noise, in order to generate appearance-conditioned frames.

Let $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}^1, \boldsymbol{\epsilon}^2, \ldots, \boldsymbol{\epsilon}^N]$ denote the noise corresponding to a video clip with N frames, $\boldsymbol{\epsilon}^i$ represents the noise added to the i^{th} frame. $\boldsymbol{z}^c$ is the latent tensor of center frame, $\boldsymbol{\epsilon}_n^i$ is the randomly sampled noise from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The training noise for our model is defined as:

$$\boldsymbol{\epsilon}^i = \lambda \boldsymbol{z}^c + \boldsymbol{\epsilon}_n^i, \tag{3}$$

where λ is the coefficient that controls the amount of the center frame.

Consequently, the diffusion process of our model can be expressed in the following form, the t-step noisy input of the diffusion model is:

$$\boldsymbol{z}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon} \tag{4}$$

where $\boldsymbol{z}_0$ is the latent tensors of an input video and $\bar{\alpha}_t$ is the same as defined in DDPM [13].

Table 1. Comparison on the zero-shot text-to-video generation performance on UCF-101[38] and MSR-VTT[51]

Methods	UCF-101[38]		MSR-VTT[51]	
	FVD ↓	IS ↑	FVD ↓	CLIPSIM ↑
Using WebVid-10M and additional data for training				
Make-A-Video [36]	367.23	33.00	-	0.3049
VideoFactory [42]	410.00	-	-	0.3005
ModelScope [41]	410.00	-	550.00	0.2930
Lavie [45]	526.30	-	-	0.2949
VidRD [9]	363.19	39.37	-	-
PYoCo [7]	355.19	**47.76**	-	**0.3204**
Using WebVid-10M only for training				
LVDM [10]	641.80	-	742.00	0.2381
CogVideo [17]	701.59	25.27	1294	0.2631
MagicVideo [55]	699.00	-	998.00	-
Video LDM [5]	550.61	33.45	-	0.2929
VideoComposer [43]	-	-	580	0.2932
VideoFusion [23]	639.90	17.49	581.00	0.2795
SimDA [50]	-	-	456.00	0.2945
Show-1 [52]	394.46	35.42	538.00	0.3072
MicroCinema (Ours)	**342.86**	37.46	**377.40**	0.2967

For training, we adhere to the stable diffusion training setting and use noise prediction with the following loss function:

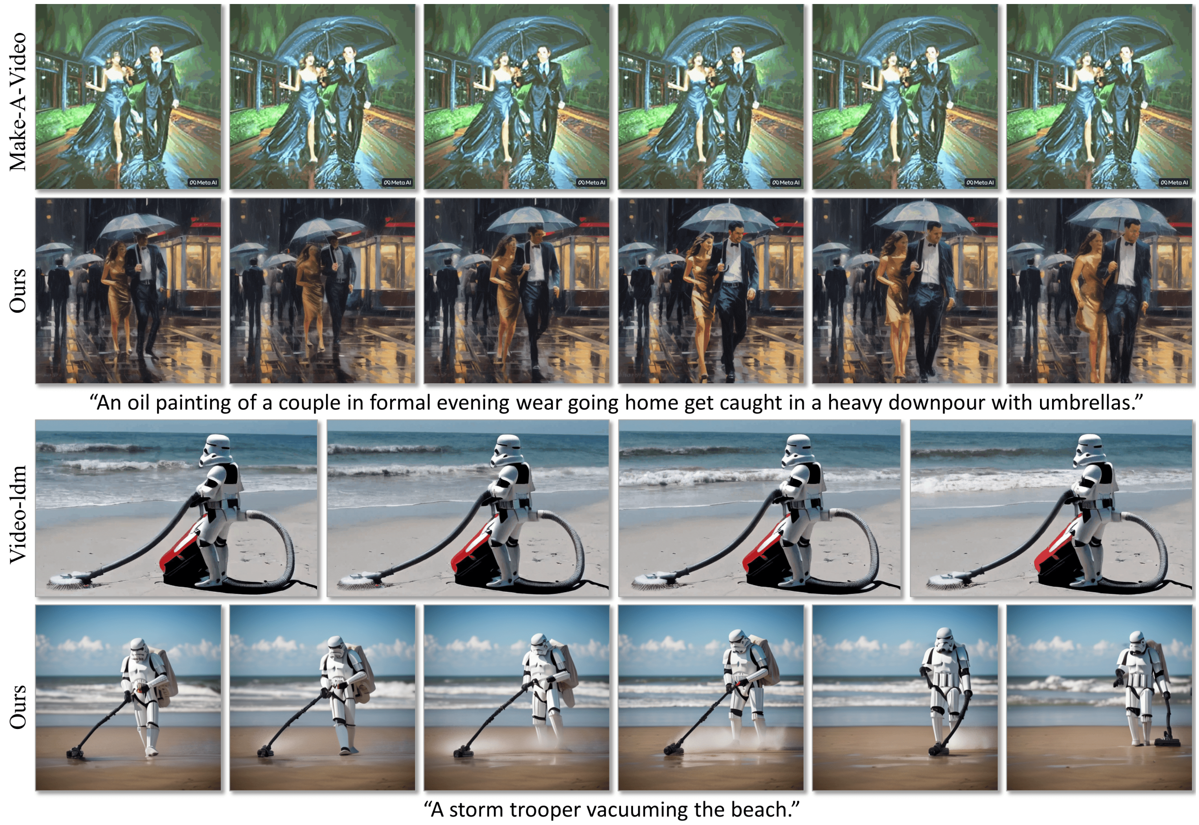
$$\mathcal{L}_{\boldsymbol{\theta}} = \mathbb{E}_{q_t(\boldsymbol{z}_0, \boldsymbol{z}_t)}\left[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{z}^c, \boldsymbol{c}) - \boldsymbol{\epsilon}\|^2\right], \tag{5}$$

where t is the time step, $\boldsymbol{z}^c$ is the reference image input, $\boldsymbol{c}$ is the text input, $\boldsymbol{z}_0$, $\boldsymbol{z}_t$ are the ground-truth video and noisy input, $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{z}^c, \boldsymbol{c})$ represents the output of the model., respectively. Our appearance noise prior employs the same inference strategy as previous methods, differing only in the initiation of noise, which aligns with our formulation. This consistency allows for the direct application of existing ODE sample algorithms. For a thorough understanding of the proofs, please refer to the supplementary materials.

4. Experiments

Datasets. MicroCinema is trained using the public WebVid-10M dataset [4], comprising ten million video-text pairs. This dataset exhibits a wide spectrum of video motions, ranging from near-static sequences to those with frequent and abrupt scene changes. Text captions are automatically sourced from alt text, resulting in some noise. Therefore, we perform a filtering process which excludes video-text pairs with a low CLIP score or with excessively high or low motions.

Evaluation metrics. The quantitative evaluations are conducted on UCF-101 [38] and MSR-VTT [51] benchmark datasets under the zero-shot setting. On UCF-101, Frechet Video Distance (FVD) [39] and Inception Score (IS) [34] are reported to validate the temporal consistency, where



"An oil painting of a couple in formal evening wear going home get caught in a heavy downpour with umbrellas."

"A storm trooper vacuuming the beach."

Figure 4. Comparison with Make-A-Video and Video LDM. Reference images generated by DALL-E 2 (top) and Midjourney (bottom). The generated videos from our model shows a clear and coherent motion.

10K or 2K video clips are generated using a sentence template of the category names. On MSR-VTT, Frechet Video Distance (FVD) [11] and CLIPSIM [36, 47] are provided to assess the quality of generated frames and the semantic correspondence, where CLIPSIM is computed by averaging the cosine similarity of CLIP embeddings [28] between generated frames and captions. We utilized captions from the MSR-VTT validation set, comprising 2.9K entries, to generate the video clips. The condition images are generated with SDXL model [32] on all evaluations unless otherwise specified.

4.1. Comparison with State-of-the-Arts

Implementation details. MicroCinema generates video from text in a two-stage process. In the first stage, we employ a SOTA T2I model SDXL [27] to generate an image according to the text. Then in the second stage, the image&text-to-video generation model is built upon the pre-trained weights of Stable Diffusion 2.1. Temporal layer is zero initialized. During training, the learning rate for the temporal modules is set to 2e-5, while the learning rate for the spatial model is 10 times smaller than that of the temporal modules. The output of the image&text-to-video model yields a video clip with a spatial resolution of 320x320, consisting of 9 frames at a rate of 2fps. The model is trained on the filtered WebVid dataset for one epoch, employing the same diffusion noise schedule as SD2.1.

Quantitative evaluation. We evaluate zero-shot text-to-video generation performance on both UCF101 and MSR-VTT. In the case of UCF101, we produce 10K samples using simple clip captions. For MSR-VTT, we generate 2.9K samples using the captions provided within the MSR-VTT dataset. Tab. 1 presents a quantitative comparison between MicroCinema and alternative text-to-video models. These models are categorized into two groups based on whether they leverage additional data beyond WebVid-10M. As data is of paramount importance to the training of video generation model, we can observe that the methods in the first group (with additional data) achieve superior overall performance compared to those in the second group. Remarkably, despite being exclusively trained on the WebVid-10M dataset, our proposed MicroCinema, with its innovative design, achieves the most outstanding performance among all methods on both datasets. It achieves the lowest FVD values of 342.86 on UCF101 and 377.40 on MSR-VTT. Notably, MicroCinema surpasses methods employing additional data and notably outperforms those relying solely on the WebVid-10M dataset by a considerable margin.

Qualitative evaluation. Fig. 4 compares the video clips generated by MicroCinema and two other methods, known

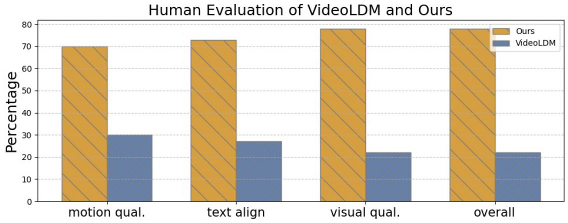


Figure 5. Human evaluation between VideoLDM and ours.

as Make-A-Video and Video LDM. Compared to the other two methods, our approach can generate noticeable and accurate motion.

Human evaluation. We Randomly selected 40 out of 54 text-video pairs from the VideoLDM, we ensured fairness by cropping appropriate region from VideoLDM videos into square shapes to match our videos and prevent information leakage. Ten university students, unfamiliar with both VideoLDM and our results, evaluated these pairs based on Motion Quality, Text Alignment, Visual Quality, and Overall Preference. Our results, depicted in the Fig. 5, conclusively show superior performance of our approach over VideoLDM across all evaluated aspects.

4.2. Ablation Studies

We conduct ablation studies to validate our design choices concerning appearance injection and shifted noise training. For efficiency purposes, we adopt several different settings from the experiments used for system comparison. First, models employing different options are trained using a 1M subset of the filtered WebVid-10M dataset. Each model undergoes training for 64K steps (equivalent to one epoch) with a batch size set at 16. Second, during inference, we directly generate 17 frames without using the TSR module. Third, for the zero-shot FVD and IS evaluation on UCF101, we uniformly select 2K samples instead of using the entire 10K test set. It's notable that while using this smaller 2K-sample test set, the absolute FVD values are higher compared to those derived from the larger 10K-sample test set for the same model.

4.2.1 Appearance Injection

In an image&text-to-video model, the most important design choice is how to inject the appearance information into the primary U-Net of the generation model.

Concatenation (Concat). A common approach in related work [5, 20] is to direct concatenation of the latent features from the reference image to the noise input of the U-Net.

Addition to Decoder (Add-to-Dec). Our approach, however, adopts an AppearNet, akin to ControlNet for structure control. In the vanilla ControlNet, embeddings from the ControlNet are added to the decoder of the U-Net. We employ a similar operation in this setting.

Addition to Encoder and Decoder (Add-to-EncDec). Considering that the reference image contains more appearance details than the structural information in ControlNet, we propose injecting appearance into both the encoder and the decoder of the U-Net. This improvement is expected to elevate generation quality through a more comprehensive integration of appearance features.

Table 2. Ablation study on UCF-101 for appearance injection methods.

Method	Zero-Shot	IS (↑)	FVD (↓)
Concat	Yes	15.83	688.92
Add-to-Dec	Yes	27.90	589.59
Add-to-EncDec	Yes	27.25	525.02
Add-to-EncDec-SPADE	Yes	29.63	508.56

Addition to Encoder and Decoder with SPADE (Add-to-EncDec-SPADE). Expanding further, we integrate the SPADE technique, commonly used in image generation models, by infusing information into the GroupNorm layers of the U-Net. This final design constitutes the core of our method, MicroCinema.

Tab. 2 presents a comparative analysis of the zero-shot FVD performance among these four design choices. The results clearly demonstrate that our final model achieves the most superior performance.

4.2.2 Appearance Noise Prior

Another key mechanism we propose for injecting appearance information into the image&text-to-video generation network is the Appearance Noise Prior. One crucial and intricate parameter within this mechanism is the proportion, denoted by λ, determining the addition of the reference image to the noise input of the diffusion model. Selecting an optimal value for λ involves balancing potential harm to the pre-trained image generation model and the advantages gained from additional information.

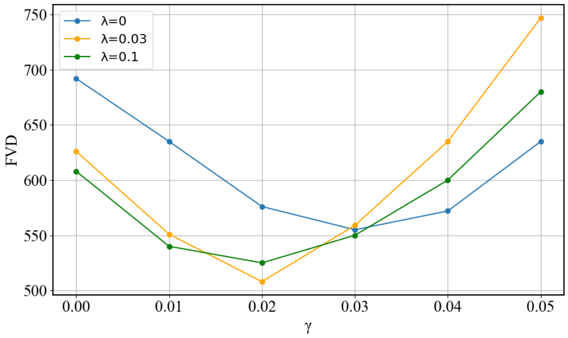


Figure 6. UCF-101 Zero shot FVD across different λ and γ.

Quantitative evaluation. This set of ablation studies aims to empirically identify the most effective parameter for use with Appearance Noise Prior. Alongside λ, which we test at values of 0 (no Appearance Noise Prior), 0.03, and 0.1. Besides, according to our formulation, an appropriate amount of appearance may also help during the inference stage.

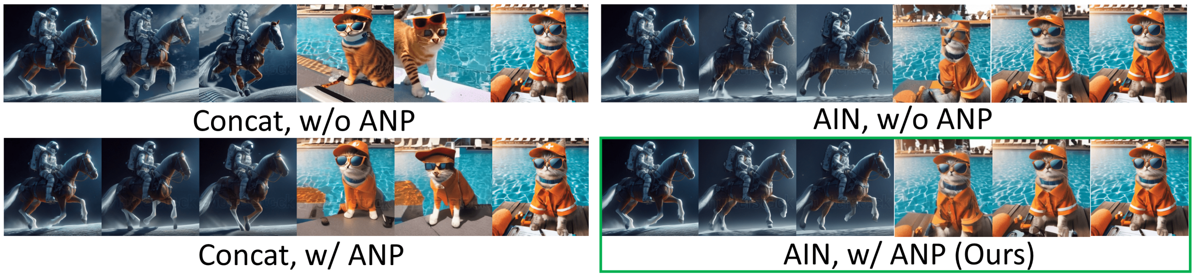


Figure 7. Qualitative ablation studies of Appearance Injection Network (AIN) and Appearance Noise Prior (ANP).

Therefore, we also explore the impact of adding extra $\gamma \boldsymbol{z}^{\mathbf{c}}$ to ϵ during the inference stage. Therefore, the sampling noise during the inference stage is $(\lambda + \gamma)\boldsymbol{z}^{\mathbf{c}} + \boldsymbol{\epsilon}_n$, where ϵ_n is sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Fig. 6 shows the FVD scores across various combinations of (λ, γ). We find that the lowest FVD score occurs when $\lambda = 0.03$ and $\gamma = 0.02$. Notably, this configuration leads to a substantial reduction in FVD compared to the baseline ($\lambda = 0$, $\gamma = 0$), dropping from 692 to 508, alongside a notable increase in IS from 18.5 to 29.6.

Human evaluation. We generated samples with or without appearance noise prior (ANP) for 20 prompts, and another 10 students are invited to the study. Their preferences, plotted in the Fig. 8, show the effect of ANP.

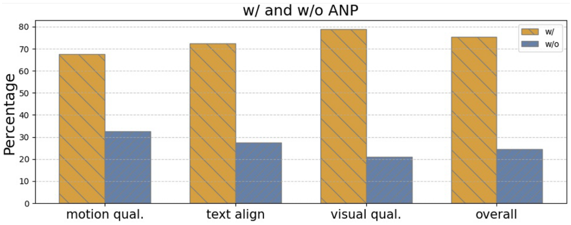


Figure 8. Human evaluation of appearance noise prior.

4.2.3 Qualitative Ablation Studies

Fig. 7 qualitatively confirms the standalone effectiveness of Appearance Injection Network and Appearance Noise Prior. Yet, their combination yields the best results.

4.3. Control in Image&Text-to-Video Model

Our image&text-to-video model relies on both a reference image and a text prompt for conditioning. Our findings emphasize that the reference image's quality profoundly influences the resultant video quality. Consequently, both the text caption and the text-to-image model used to generate the reference image significantly impact the system's performance. We simplify our experiments by using the base image&text-to-video model without using the temporal super-resolution component. In this setup, we adopt the resulting model to generate 17 frames with 10K samples on UCF101 for evaluating IS and FVD.

Tab. 3 illustrates the influence of various prompts on the model's generated outputs. We utilize the state-of-the-art SDXL model for text-to-image generation. Within the table, "simple" denotes a straightforward prompt created by connecting "a video of" with the motion tag, while "LLaVA-1.5" signifies a generated caption via the LLaVA-1.5 model [21, 22] using the key frame as input. Results indicate that a well-crafted prompt correlates with higher-quality videos generated by the model. Moreover, we assess the impact of employing different Text-to-Image (T2I) models. Tab. 4 underscores the substantial influence of T2I models on the FVD and IS of the generated videos. Notably, the design of MicroCinema affords us the flexibility to integrate various T2I models for generating the first-stage reference image, with potential performance enhancements stemming from advancements in text-to-image models.

Table 3. Evaluation on UCF-101 using different text prompts. SDXL is used as the first stage model.

Method	Prompt	IS (↑)	FVD (↓)
MicroCinema	Simple	29.79	374.05
MicroCinema	LLaVA-1.5	32.07	336.40

Table 4. Evaluation on UCF-101 using different text-to-image models. Prompts generated by LLaVA-1.5 are utilized.

Method	Frist Stage Model	IS (↑)	FVD (↓)
MicroCinema	SD-2.1	31.25	412.53
MicroCinema	SDXL	32.07	336.40

5. Conclusion

We presented MicroCinema, an innovative text-to-video generation approach that employs the Divide-and-Conquer paradigm to tackle two key challenges in video synthesis: appearance generation and motion modeling. Our strategy employs a two-stage pipeline, utilizing any existing text-to-image generator for initial image generation and subsequently introducing a dedicated image&text-to-video framework designed to focus on motion modeling. To improve motion capture, we propose an Appearance Injection Network structure, complemented by an appearance-aware noise prior. Experimental results showcase MicroCinema's superiority, achieving a state-of-the-art zero-shot Frechet Video Distance (FVD) of **342.86** on UCF101 and **377.40** on MSR-VTT. We anticipate our research will inspire future advancements in this direction.

References

[1] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023. 3

[2] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017. 2

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 2

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 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*, pages 1728–1738, 2021. 2, 5

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 3, 4, 5, 7

[6] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2015. 2

[7] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 2, 3, 5

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[9] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023. 5

[10] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2023. 5

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 30, 2017. 6

[12] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 2

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 5

[14] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 3

[15] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2, 3

[16] Susung Hong, Junyoung Seo, Sunghwan Hong, Heeseong Shin, and Seungryong Kim. Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*, 2023. 3

[17] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3, 5

[18] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 3

[19] Ariel Lapid, Idan Achituve, Lior Bracha, and Ethan Fetaya. Gd-vdm: Generated depth for better diffusion-based video generation. *arXiv preprint arXiv:2306.11173*, 2023. 3

[20] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 3, 7

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 8

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 8

[23] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 5

[24] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[25] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 4

[26] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 2

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 6

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6

[29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[31] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 6

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 2, 5

[35] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[36] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 3, 5, 6

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[38] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 5

[39] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. *ICLR Workshop: Deep Generative Models for Highly Structured Data*, 2019. 5

[40] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 2

[41] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 5

[42] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 5

[43] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 5

[44] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022. 3

[45] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 5

[46] Yaohui Wang, Xin Ma, Xinyuan Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. Leo: Generative latent image animator for human video synthesis. *arXiv preprint arXiv:2305.03989*, 2023. 3

[47] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 6

[48] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4

[49] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943*, 2023. 3

[50] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. *arXiv preprint arXiv:2308.09710*, 2023. 5

[51] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 2, 5

[52] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 5

[53] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 3

[54] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 3

[55] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video

generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 5