

Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance¹

Jinbo Xing¹ Menghan Xia^{2,*} Yuxin Liu¹ Yuechen Zhang¹ Yong Zhang² Yingqing He³ Hanyuan Liu¹
Haoxin Chen² Xiaodong Cun² Xintao Wang² Ying Shan² Tien-Tsin Wong¹
¹CUHK ²Tencent AI Lab ³HKUST

*Corresponding Author. Project page: <https://doubiu.github.io/projects/Make-Your-Video>



Figure 1. Given the text descriptions and motion structure as guidance, our model can generate temporally coherent videos adhering to the guidance intentions. By building structural guidance from distinct sources, we show the video generation results in different applications, including (top) real-life scene setup to video, (middle) dynamic 3D scene modeling to video, and (bottom) video re-rendering.

Abstract

Creating a vivid video from the event or scenario in our imagination is a truly fascinating experience. Recent advancements in text-to-video synthesis have unveiled the potential to achieve this with prompts only. While text is convenient in conveying the overall scene context, it may be insufficient to control precisely. In this paper, we explore customized video generation by utilizing text as context description and motion structure (e.g. frame-wise depth) as concrete guidance. Our method, dubbed Make-Your-Video, involves joint-conditional video generation using a Latent Diffusion Model that is pre-trained for still image synthesis and then promoted for video generation with the introduc-

tion of temporal modules. This two-stage learning scheme not only reduces the computing resources required, but also improves the performance by transferring the rich concepts available in image datasets solely into video generation. Moreover, we use a simple yet effective causal attention mask strategy to enable longer video synthesis, which mitigates the potential quality degradation effectively. Experimental results show the superiority of our method over existing baselines, particularly in terms of temporal coherence and fidelity to users' guidance. In addition, our model enables several intriguing applications that demonstrate potential for practical usage.

1. Introduction

As a widely embraced digital medium, videos are highly regarded for their ability to deliver vibrant and immersive visual experiences. Capturing real-world events in video has become effortless with the widespread availability of smartphones and digital cameras. However, when it comes to creating a video to express the idea aesthetically, the process becomes considerably more challenging and costly, which usually requires professional expertise in computer graphics, modeling, and animation production. Fortunately, recent advancements in text-to-video [11, 41] shed light on the possibility of simplifying this process as textual prompts alone. Although text is recognized as a standard and versatile description tool, we argue that it excels primarily in conveying abstract global context, while it may be less effective in providing precise and detailed control. This motivates us to explore customized video generation by utilizing text as context description and motion structure as concrete guidance.

Specifically, we choose frame-wise depth maps to represent the motion structure, as they are 3D-aware 2D data that align effectively with the task of video generation. In our approach, the structural guidance can be quite rough, so as to allow non-professionals to easily prepare it. This design offers the flexibility for the generative model to produce plausible content without depending on intricately crafted input. For instance, a scene setup using office common items can be used to guide the generation of a photorealistic outdoor landscape (Figure 1(top)). With 3D modeling software, the physical objects can be replaced with simple geometric elements or any accessible 3D assets (Figure 1(middle)). Naturally, another alternative is to make use of the estimated depth from existing videos (Figure 1(bottom))). So, the combination of textual and structural guidance provides users with both flexibility and controllability to customize their videos as intended.

To achieve this, we formulate the conditional video generation using a Latent Diffusion Model (LDM) [36] that adopts a diffusion model in a compressed lower-dimensional latent space to reduce the computational expenses. For training an open-world video generation model, we propose to separate the training of spatial modules (for image synthesis) and temporal modules (for temporal coherence). This design is based on two primary considerations: (i) training the model components separately eases the computational resource requirements, which is especially crucial for resource-intensive tasks; (ii) as image datasets encompass a much enriched variety of concepts than the existing video datasets, hence, pre-training the model for image synthesis helps to inherit the diverse visual concepts and transfers them to video generation. A major challenge is to achieve temporal coherence. Specifically, using a pre-trained image LDM, we maintain them as

the frozen spatial blocks and introduce the temporal blocks dedicated to learning inter-frame coherence over the video dataset. Notably, we combine the temporal convolutions with the additional spatial convolutions, which improves the temporal stability by increasing the adaptability to the pre-trained modules. We also adopt a simple yet effective causal attention mask strategy to enable longer (i.e., $4 \times$ the training length) video synthesis and it mitigates the potential quality degradation significantly.

Both qualitative and quantitative evaluations evidence the superiority of our proposed method over existing baselines, particularly in terms of temporal coherence and fidelity to users' guidance. Ablation studies confirm the effectiveness of our proposed designs, which play a crucial role in the performance of our method. Additionally, we showcased several intriguing applications facilitated by our approach, and the results indicate the potential for practical scenarios.

Our contributions are summarized as below:

- We present an efficient approach for customized video generation by introducing textual and structural guidance. Our method achieves top performance in controllable text-to-video generation both quantitatively and qualitatively.
- We propose a mechanism to leverage pre-trained image LDMs for video generation, which inherits the rich visual concepts while achieving a decent temporal coherence.
- We introduce a temporal masking mechanism to allow longer video synthesis while alleviating the quality degradation.

2. Related Work

2.1. Diffusion Models for Text-to-Image (T2I) Synthesis

Diffusion models [12, 42, 44] (DMs) have recently shown unprecedented generative semantic and compositional power, attracting attention from both academia and industry. By absorbing the text embedding from CLIP [31] or T5 [32], they have been successfully adopted for text-to-image synthesis [2, 7, 28, 33, 36, 38]. GLIDE [28] introduces classifier-free guidance [13] to the text-conditioned image synthesis DMs, improving image quality in both photorealism and text-image alignment, which are further boosted by using CLIP [31] feature space in DALL·E 2. Moreover, a line of works [20, 27, 51–53] improves the controllability of T2I by introducing additional conditional inputs, *e.g.*, pose, depth, and normal map. Since DMs generally require iterative denoising processes through a large U-Net, the training becomes computationally expensive. To ad-

dress this, cascaded (Imagen [38]) and latent diffusion models (LDMs [36]) have been proposed. Specifically, Imagen adopts cascaded diffusion models in pixel space to generate high-definition videos, while LDMs first compress the image data using an autoencoder and learn the DMs on the resultant latent space to improve efficiency. To inherit the diverse visual concepts and reduce the training cost, our method builds upon LDMs by introducing video awareness through additional learnable temporal modules into the architecture and training on text-video data, while keeping the original weights of LDMs frozen.

2.2. Diffusion Models for Text-to-Video (T2V) Synthesis

Although there have been significant advancements in T2I, text-to-video generation (T2V) is still lagging behind due to the scarcity of large-scale high-quality paired text-video data, the inherent complexity of modeling temporal consistency, and resource-intensive training. As a pioneering work, Video Diffusion Model [14] models low-resolution videos with DMs using a space-time factorized U-Net in pixel space and trains jointly on image and video data. To generate high-definition videos, Imagen-Video [11] proposes effective cascaded diffusion models and the v-prediction parameterization method.

To reduce the training cost, many subsequent studies, including Make-A-Video [41], MagicVideo [54] and LVDM [9], transfer T2I knowledge to T2V generation by initiating from pre-trained T2I models and fine-tuning the entire model. Differently, Khachatryan *et al.* [19] present a training-free transfer by leveraging pre-trained T2I models with manual pseudo motion dynamics to generate short videos. However, the generated videos suffer from low quality and inconsistency. Besides adopting the pre-trained T2I, Gen-1 [6] and FollowYourPose [23] propose to control the structure and motion dynamics of synthesized videos by depth and pose, respectively. With the same conditions as us, Gen-1, however, trains the entire model, which can be both time- and resource-consuming, potentially leading to a degradation of the inherited rich visual concepts. As a concurrent work, Video-LDM [4] shares a similar motivation with ours, *i.e.*, extending the image LDMs to video generators by introducing temporal layers and keeping the original weights frozen. However, solely using temporal layers may not be sufficient for adapting LDMs to video generators.

As for longer video synthesis, interpolation [4, 9, 41, 54] and prediction [4, 8] strategies are commonly adopted in existing diffusion-based T2V approaches. Unlike prediction, interpolation does not increase the physical time span of synthesized videos but only makes enhance their smoothness. However, the current prediction mechanism in video diffusion models is still limited in the curated domain, *e.g.*, driving [4] and indoor scene [8]. Our work aims to effi-

ciently and effectively adapt a pre-trained T2I model to a joint text-structure-guided video generator, and investigate general video prediction mechanisms for longer video synthesis.

2.3. Text-driven Video Editing

In recent studies, DMs have demonstrated their efficacy in image editing tasks, as evidenced by several works [5, 10, 18, 24, 46]. However, their application to video editing on individual frames can result in temporal inconsistency issues. To address this, Text2LIVE [3] combines Layered Neural Atlases [17] and the proposed text-driven image editing method, allowing texture-based video editing but struggling to accurately reflect the intended edits. To improve the video quality, recent diffusion-based video editing methods rely on either pre-trained large-scale video diffusion models [26], which are usually inaccessible and hard to reproduce due to the unaffordable training, or the inversion [25, 43] followed by attention manipulation mechanism [21, 30, 49] using pre-trained T2I model, rendering tricky prompt engineering or manual hyper-parameter tuning process. Although both our conditional video generator and video editing methods can edit video content, a notable distinction is their reliance on the original video (*e.g.*, for inversion purposes), whereas our model does not necessitate the source video as input.

3. Method

The goal of this work is to study controllable text-to-video synthesis so that the generated video could align with the users' intention faithfully. To achieve this, we propose a conditional video generation model that takes text prompts and frame-wise depths as conditional input. The text prompt describes the video appearance and depth sequence specifies the overall motion structure.

In the following, we start with a brief introduction to DMs and LDMs (Section 3.1), as the preliminary knowledge to our proposed controllable text-to-video generation model (Section 3.2). At last, we discuss on a temporal masking mechanism to facilitate longer video generation (Section 3.3).

3.1. Preliminaries

Diffusion models (DMs) are probabilistic models designed to learn a target data distribution $p_{\text{data}}(\mathbf{x})$ by gradually denoising a normally distributed variable. This denoising process corresponds to learning the reverse process of a fixed Markov Chain of length T with denoising score matching [16, 22, 44]. The most successful models in the image synthesis field rely on a reweighted variant of the variational lower bound on $p_{\text{data}}(\mathbf{x})$. These models can be interpreted as an equally weighted sequence of denoising autoencoders $\epsilon_\theta(\mathbf{x}_t, t)$, parameterized with learnable parame-

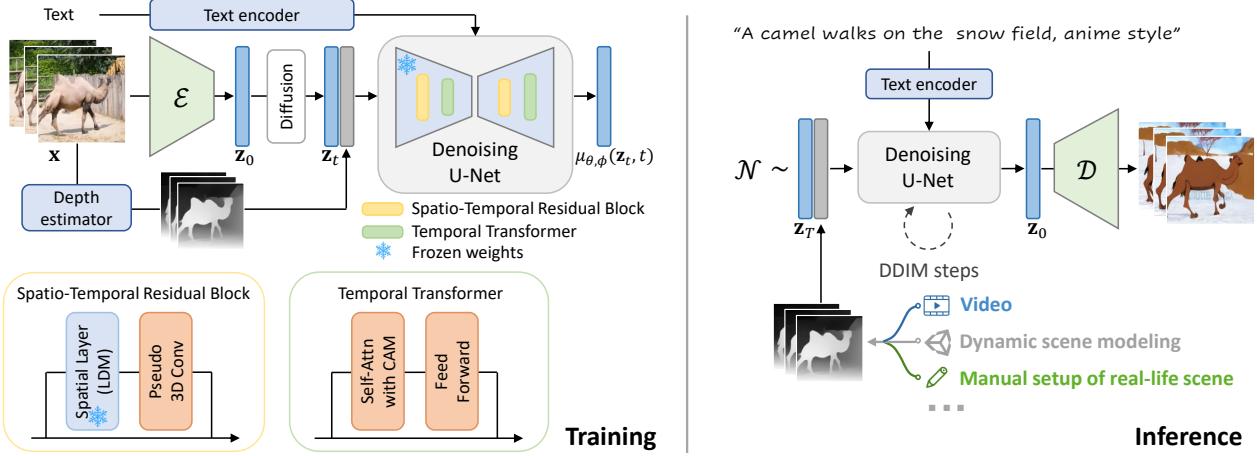


Figure 2. Flowchart of the proposed method. During training (left), the input video \mathbf{x} is first encoded into latent feature \mathbf{z}_0 with a fixed pre-trained encoder \mathcal{E} and diffused to \mathbf{z}_t . Meanwhile, the depth sequence will be extracted with the off-the-shelf depth estimator MiDas and concatenated with \mathbf{z}_t , and the text is encoded by a frozen OpenCLIP text encoder. Then the model learns to reverse the diffusion process conditioned on the depth and text prompt. As for inference (right), videos can be generated by recurrently denoising a random tensor sampled from normal distribution, under the guidance of text prompt and frame-wise depth obtained in multiple ways.

ters θ ; $t = 1 \dots T$, which are trained to predict a denoised variant of their input \mathbf{x}_t , where \mathbf{x}_t is a noisy version of the input $\mathbf{x} \sim p_{\text{data}}$. The corresponding denoising objective is

$$\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t; \mathbf{c}, t)\|_2^2], \quad (1)$$

where \mathbf{c} is optional conditioning information (e.g. text prompt), t denotes a timestep uniformly sampled from $\{1, \dots, T\}$, and ϵ is the noise tensor used during the diffusion from \mathbf{x}_0 to \mathbf{x}_t . The neural backbone implementation of $\epsilon_\theta(\cdot; \mathbf{c}, t)$ is generally a 2D U-Net [37] with cross-attention conditioning mechanisms.

Latent diffusion models (LDMs) [36] are proposed to improve the computational and memory efficiency over a learned compact latent space instead of the original pixel space. It is realized through perceptual compression with an autoencoder \mathcal{E} and \mathcal{D} for efficient and spatially lower-dimensional feature representations. The autoencoder is defined to reconstruct inputs \mathbf{x} , such that $\hat{\mathbf{x}} = \mathcal{D}(\mathcal{E}(\mathbf{x})) \approx \mathbf{x}$. A DM can then be trained in the compressed latent space and turned into a LDM. The corresponding objective is similar to Eq. 1, except for replacing \mathbf{x} with its latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$.

3.2. Adapting LDMs for Conditional T2V Generation

We employ an LDM to formulate our conditional video generation task, which involves synthesizing content for each frame while maintaining their temporal coherence to produce plausible dynamics. Our key insight is to harness the power of pre-trained conditional T2I LDMs as a language-visual generative prior for conditional video synthesis. The main challenges are two-fold: (i) although the

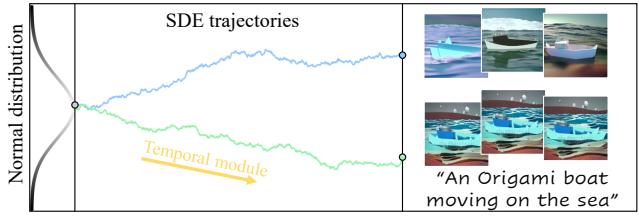


Figure 3. Concept depiction of the effect of temporal modules ϵ_ϕ . Given a batch of noises, ϵ_ϕ pushes the SDE trajectories toward target data distributions with cross-frame coherence. We draw a single trajectory for a batch for conciseness.

pre-trained image LDM can synthesize high-quality individual frames, it is not ready for generating temporally consistent video frames; and (ii) there is a shortage of large-scale text-video datasets with rich concept coverage and high quality when compared to image datasets like LAION [39]. To address the first issue, we promote the original depth-conditioned LDM (CLDM) to video generators by introducing additional temporal modules. In principle, these modules push the reverse diffusion process of a pre-trained image CLDM toward a certain required subspace, which is conceptually shown in Figure 3. As for the second one, we freeze the weights of the pre-trained CLDM to lock the learned image prior and only fine-tune the temporal modules over a video dataset. The proposed framework is illustrated in Figure 2.

Specifically, during training, the input video \mathbf{x} is first encoded into the latent feature $\mathbf{z}_0 \in \mathbb{R}^{L \times C \times H' \times W'}$ in a frame-wise manner using pre-trained encoder \mathcal{E} , where C is the number of latent channel dimensions, H' and W' are

the latent spatial size, *i.e.*, height and width, respectively. To inject structural guidance into denoising, we concatenate the frame-wise depth s extracted from the input video with an off-the-shelf depth estimator MiDAS DPT-Hybrid [34] with \mathbf{z}_t that is diffused from latent feature \mathbf{z}_0 . The textual guidance is introduced through cross-attention layers in the OpenCLIP text embedding space. The original image CLDM layers process the video data as a batch of independent input images by shifting the temporal dimension to the batch dimension for frame-wise processing. Accordingly, the batched feature output will be shifted back to video before being fed to the temporal modules.

We implement two types of temporal modules ϵ_ϕ , namely, Spatio-Temporal Residual Block (STRB) and Temporal Transformer (TT), shown at the bottom of Figure 2. STRB is an extension of the original residual block containing spatial layers only. To make it capable of capturing temporal priors in videos, we introduce additional pseudo 3D convolution layers, *i.e.*, 2D spatial conv followed by 1D temporal conv, based on two considerations: (i) there is a domain gap between the modeled distribution from text-image dataset and target distribution in text-video data, as least in content; and (ii) the additional spatial layers could assist the temporal layers in learning motion dynamics by increasing the adaptability to those pre-trained spatial modules, which is evidenced by the experiments in Section 4.3. The Temporal Transformer is located behind the original spatial transformer and its design is to leverage the property of temporal self-similarity to learn the inherent motion dynamics along the temporal axis in video data. Concretely, it consists of temporal self-attention modules with learnable positional embeddings [40], and a feed-forward layer [48].

Then our controllable T2V framework with denoiser $\epsilon_{\theta, \phi}$ can be trained with a similar setting to the underlying CLDM. It is worth noting that the original weights ϵ_θ are frozen and only the added temporal modules ϵ_ϕ are learned via the optimization objective:

$$\arg \min \mathbb{E}_{\mathcal{E}(\mathbf{x}), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} [\|\epsilon - \epsilon_{\theta, \phi}(\mathbf{z}_t; \mathbf{c}, \mathbf{s}, t)\|_2^2], \quad (2)$$

where \mathbf{z}_t indicates the diffused latent $\mathbf{z} = \mathcal{E}(\mathbf{x})$ with noise level t .

During inference, as depicted in Figure 2 (right), the depth s obtained from multiple sources for structure control and the text prompt \mathbf{c} describing the target appearance serve as conditions for the reverse diffusion process. This process starts with randomly sampled noise \mathbf{z}_T and the denoised latent \mathbf{z}_0 is converted into a video in pixel space using the pre-trained decoder \mathcal{D} .

3.3. Temporal Masking for Longer Video Synthesis

Our conditional video LDM can generate satisfactory videos with the same number of frames (*i.e.*, 16 frames)



Figure 4. Illustration of the causal attention mask during training & inference.

as in the training phase. However, when using it to generate longer videos during inference, we observe significant quality degradation. The possible reason is that the temporal self-attention is conducted in an N-to-N manner and the learned parameters are well-fitted to process a fixed length of tokens. According to this conjecture, longer token sequences tend to disturb each other because of the confused attention across frames.

To alleviate this problem, we propose to introduce a temporal masking mechanism so that the learned temporal attention module can better adapt to the cases involved in longer video synthesis. As shown in Figure 4, we adopt the causal attention mask (CAM) strategy to achieve this. The temporal attention \mathbf{F}_t of an input feature \mathbf{z}_t is calculated via:

$$\mathbf{F}_t = \text{Attention}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t) = \text{softmax}\left(\frac{\mathbf{Q}_t \mathbf{K}_t^\top}{\sqrt{d}} + \mathbf{M}\right) \mathbf{V}_t, \quad (3)$$

where \mathbf{Q}_t , \mathbf{K}_t , \mathbf{V}_t are linearly projected features from \mathbf{z}_t , d denotes the head dimension, and \mathbf{M} is a lower triangular matrix ($M_{i,j} = 0$ if $i > j$ else $-\infty$) during training. For longer video synthesis during inference, the mask is modified to ensure the present token is only affected by the previous L_M tokens, *i.e.*, maximum temporal receptive field. With the help of CAM, the self-attention layers can be aware of different lengths of tokens, making the causal receptive field adjustable. It can thus effectively mitigate the quality degradation and temporal inconsistency problem for longer video synthesis. Still, larger L_M leads to better cross-frame coherence while hinders the video quality, and vice versa.

Moreover, the CAM is an explicit way to inject video prior (*i.e.*, motions in the video data are directional) into the adapted image LDM to learn temporal coherence, which could benefit even when generating short sequences, as evidenced in Section 4.3. Thus we also adopt the causal attention mask in all experiments.

Table 1. Quantitative comparisons with state-of-the-art conditional video generation methods on UCF-101 for the zero-shot setting.

Method	Condition	FVD ↓	KVD ↓
CogVideo (Chinese)	Text	751.34	-
CogVideo (English)	Text	701.59	-
Magic Video	Text	699.00	-
Make-A-Video	Text	367.23	-
Video LDM	Text	550.61	-
T2V-zero+CtrlNet	Text+Depth	951.38	115.55
LVDM _{Ext} +Adapter	Text+Depth	537.85	85.47
Ours w/o CAM	Text+Depth	390.63	36.57
Ours	Text+Depth	330.49	29.52

4. Experiments

4.1. Implementation Details

Our development is based on depth-conditioned Latent Diffusion Models [36] (a.k.a Stable-Diffusion-Depth) implemented with PyTorch and the public pre-trained weights*. We train the newly introduced layers with 50K steps on the learning rate 1×10^{-4} and valid mini-batch size 512 with DeepSpeed [35]. At inference, we use DDIM sampler [43] with classifier-free guidance [13] in all experiments.

Datasets. We use the WebVid-10M [1] dataset to turn the depth-conditioned LDM into a controllable text-to-video generator. WebVid-10M consists of 10.7 million video-caption pairs with a total of 52K video hours and is diverse and rich in content. During training, we sample 16 frames with a frame stride of 4 (assuming 30 FPS) and a resolution of 256 x 256 from input videos.

4.2. Evaluation on Video Generation

Joint text-structure-conditioned video synthesis is a nascent area of computer vision and graphics, thus we find a limited number of publicly available research works to compare against. The extension version of LVDM [9] and T2V-zero [19] are general text-to-video methods but capable of generating videos with additional conditions supported by ControlNet [53] or Adapter [27], and we denote them as LVDM_{Ext}+Depth Adapter and T2V-zero+Depth CtrlNet, respectively. Meanwhile, we benchmark against pure text-to-video synthesis methods, including CogVideo [15], MagicVideo [54], Make-A-Video [41], and Video LDM [4]. Since there is no structure control for these approaches, we include them here for reference and to examine the performance differences concerning structural guidance in text-

*<https://huggingface.co/stabilityai/stable-diffusion-2-depth>

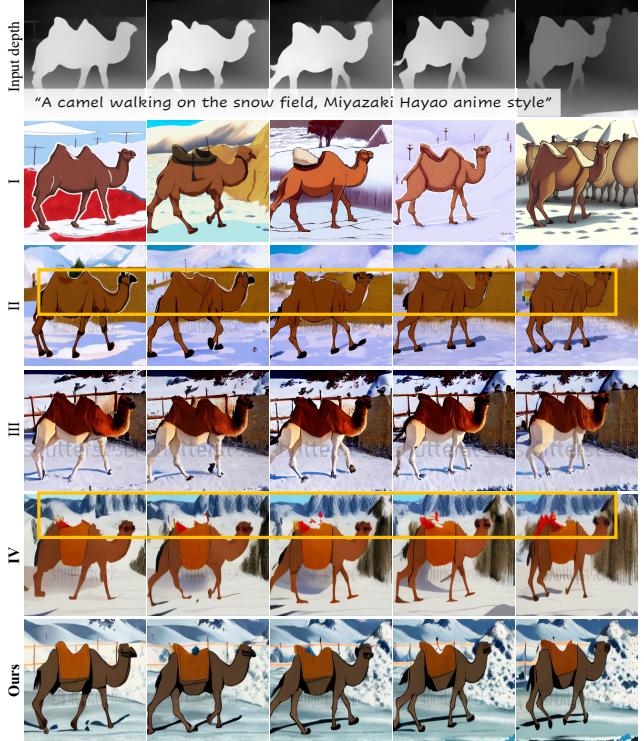


Figure 5. Visual comparisons of the videos synthesized by different variants of our approach.

to-video synthesis.

To evaluate the performance of video generation, we report the commonly-used Fréchet Video Distance (FVD) [47] and Kernel Video Distance (KVD) [47], which evaluate video quality by measuring the feature-level similarity between synthesized and real videos based on the Fréchet distance and kernel methods, respectively. Specifically, they are computed by comparing 2K model samples (16 frames) with samples from the common evaluation [4, 54] dataset UCF-101 [45]. Following [4], we directly use UCF class names as text conditioning.

We evaluate in the zero-shot setting and tabulate the quantitative performance in Table 1. According to the results, our method significantly outperforms all baselines with lower FVD and KVD. It is worth noting that although Make-A-Video with text condition only achieves an FVD value close to ours, they train on an additional extremely-large-scale dataset containing 100M text-video pairs, *i.e.*, HD-VILA-100M [50]. The superiority of our method indicates the effectiveness of the proposed video generation framework with textual and structural guidance and the adapting strategy that turns image LDMs into video generators. The qualitative comparison is made in the context of several application scenarios, as shown in Figure 7 and Figure 8.

Table 2. Ablation study of different adapting strategies on UCF-101 for zero-shot setting. (Note: TT=Temporal Transformer, TC=1D Temporal Conv, P3D=Pseudo 3D Conv)

Variant	Fine-tuned para.	FVD ↓	KVD ↓
I. SD-Depth	None	1422.30	316.25
II. w/ TT	TT	803.24	89.26
III. w/ TT	Full U-Net	500.96	77.72
IV. w/ TT+TC	TT+TC	443.63	40.86
V. w/ TT+P3D (Ours)	TT+P3D	330.49	29.52

4.3. Ablation Studies

We study several key designs of our proposed method in this section, including the adapting strategy and causal attention mask.

Adapting strategy. To study the effectiveness and superiority of our adapting strategy, we construct several baselines: **(I.)**: SD-Depth, the pre-trained image LDM that we build upon, **(II.)**: adding Temporal Transformer (TT) to baseline **(I.)** and fine-tuning this module, **(III.)**: the same architecture as **(II.)** but fine-tune the entire model, **(IV.)**: adding both TT and 1D Temporal Convolutions (TC) to **(I.)** and fine-tuning these two modules, and **(V.)**: introducing both TT and Pseudo 3D modules to **(I.)** and fine-tuning them, which is our full method. The quantitative comparison is shown in Table 2. By comparing baselines **(I.)**, **(II.)**, and **(IV.)**, we can observe the improved performance in terms of both FVD and KVD by introducing more temporal modules that increase the temporal modeling capability. The comparison between **(V.)** and **(IV.)** highlights the benefits of incorporating additional spatial layers with TC, which enhances the adaptability between the newly introduced temporal modules and the fixed spatial modules. It is worth noting that although fine-tuning the entire model **(III.)** improves the performance quantitatively over its counterpart **(II.)**, it cause a severe concept forgetting issue, as evidenced by Figure 5 where **(III.)** fails to reflect the ‘anime style’. The visual comparison also tells the superiority of our full method in terms of both temporal coherence and conformity to the text prompt.

Causal attention mask. We further investigate the effectiveness of causal attention mask (CAM). We construct a baseline without CAM and train it with the same sequence length as our full method (16 frames). As shown in Table 1 (bottom), CAM can boost the performance of conditional video generation in a 16-frame setting, owing to the introduced directional video motion prior. We directly apply the trained models for longer video synthesis (64 frames). The qualitative comparison is presented in Figure 6, wherein our baseline variant suffers from significant quality degradation and temporal inconsistency (more evident in the supplementary video) due to the mismatched training and infer-

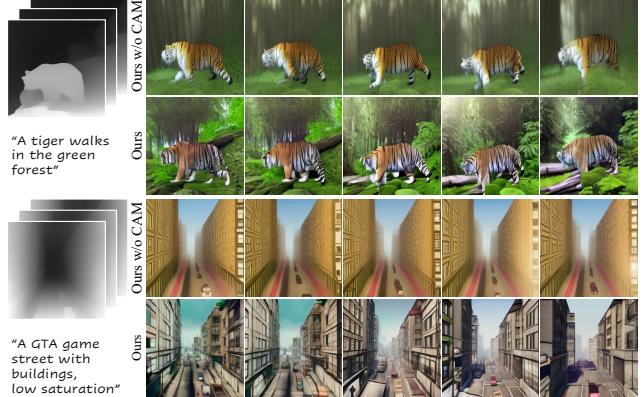


Figure 6. Visual comparisons of longer video synthesis (64 frames) produced by our baseline variant (w/o CAM) and our method (w/ CAM). Each frame is selected with a stride of 16.

ence setting. In contrast, our method could produce plausible longer videos with greater detail and improved cross-frame coherence. Readers are recommended to check our supplementary video for better comparison.

5. Applications

5.1. Video Creation

Our approach allows customized video creation guided by rough motion structures. So, it is feasible to capture a video of manually constructed miniature setup and use it as structural control for text-to-video synthesis. Some examples are illustrated in Figure 1 (top) and Figure 7 (top). In comparison to other baseline methods, our approach is capable of generating high-fidelity, temporally consistent videos that closely adhere to the target textual descriptions and scene structure. In contrast, T2V-zero+CtrlNet mainly suffers from inconsistency due to its weak temporal constraint, and LVDM_{Ext}+Adapter tends to cause lower visual quality that inherits from the base text-to-video model. One can also construct the scene structure with 3D modeling software, *e.g.*, Unity, as shown in Figure 1 (middle) and Figure 7 (bottom). For LVDM_{Ext}+Adapter, apart from the aforementioned problems, it also fails to synthesize stylized videos, *e.g.*, ‘2D cartoon’ and ‘Chinese ink wash’ at the bottom-left corner of Figure 7. It is worth noting that our method can achieve both decent text alignment and cross-frame coherence.

5.2. Video Re-rendering

Video re-rendering here means changing the video appearance based on the text prompts while still preserving its motion structure. In addition to LVDM_{Ext}+Depth Adapter and T2V-zero+Depth CtrlNet, we also compare against the depth-conditioned image LDM (*i.e.*, SD-Depth)

Table 3. Quantitative comparisons for video re-rendering.

Metric	SD-Depth	T2V-zero +CtrlNet	LVDM _{Ext} +Adapter	Tune-A-Video	Ours
Temp. \uparrow	0.8624	0.9109	0.9613	0.9659	0.9698
Prompt \uparrow	0.3695	0.3379	0.3372	0.3473	0.3547

that is used by our model as spatial modules, and a video editing method, Tune-A-Video [49] combined with DDIM inversion [43], which requires per-video optimization and the original RGB video for inversion operation. Following previous works [3, 6, 49], we use videos from DAVIS [29] and other in-the-wild videos, from which 11 representative videos with manually designed text prompts are utilized for evaluation.

Quantitative evaluation. We measure temporal consistency and prompt conformity for performance comparison. Following [6], temporal consistency is calculated as the average cosine similarity between consecutive frame embeddings of CLIP image encoder, while prompt consistency is calculated as the average cosine similarity between text and image CLIP embeddings across all frames. Table 3 shows the results of each model, where ‘Temp.’ and ‘Prompt’ indicate the temporal coherence and prompt conformity, respectively. Our model outperforms the baseline models in temporal consistency and except SD-Depth in prompt conformity. Anyhow, as an image synthesis model, SD-Depth fails to synthesize consistent video frames. Besides, although achieves the second-best temporal consistency performance, Tune-A-Video suffers from overfitting to the original video, as evidenced by the results in Figure 8.

Qualitative evaluation. We visually compare our method with other competitors in Figure 8. In the ‘camel’ case, the videos produced by our method are temporally coherent and well-align with the text description, *i.e.*, ‘anime style’. In contrast, SD-Depth cannot produce consistent video frames; Tune-A-Video, LVDM_{Ext}+Adapter, and T2V-zero+CtrlNet suffer from low text-video conformity, visual concept forgetting, and structure deviations respectively. In ‘waterfall’ case, our method shows similar superiority in the comprehensive quality including video quality, structure preservation, text-video conformity, and temporal coherence. Readers are recommended to check our supplementary video for better comparison.

User study. The human perception system is still the most reliable measure for video generation tasks. We conduct a user study, where 32 participants with good vision ability complete the evaluation successfully. For each example, the participants are first shown with the input video and target prompt, followed by five randomly ordered videos re-rendered by different methods. Then, they are asked to rank the results in terms of temporal coherence,

Table 4. User study statistics of *average ranking* \downarrow and preference rate \uparrow .

Property	SD-Depth	T2V-zero +CtrlNet	LVDM _{Ext} +Adapter	Tune-A-Video	Ours
Temporal coherence	3.95 6.53%	3.40 6.53%	2.84 12.78%	2.54 33.52%	2.27 40.62%
Structure& Text align.	3.54 13.07%	3.01 15.34%	3.10 12.50%	2.89 20.17%	2.45 38.92%
Frame quality	3.45 13.64%	3.09 13.07%	3.07 14.20%	2.84 24.15%	2.55 34.94%

text & structure guidance conformity, and frame quality: {1: the best, 2: the second-best, ..., 5: the worst}. We analyze the collected evaluation result in two aspects: (i) average ranking: the average ranking score of each method according to the rank-score table, and (ii) preference rate: the percentage of being selected as the best. The statistics are tabulated in Table 4. Our method earns the best ranking scores and preference rates in all three aspects.

6. Limitation

Our method also has certain limitations. Firstly, our customized video generation model supports no precise control over visual appearance. For instance, it is intractable to synthesize videos featuring a specific individual or object. This relates to the concept customization technique, which has been extensively studied in T2I synthesis but remains less explored in T2V synthesis. Additionally, our controllable text-to-video model explicitly demands frame-wise depth guidance, which can be overly intensive and may increase usage challenges in certain situations. A more efficient formulation could be supplying the model only with sparse keyframe guidance, which could help broaden the scope of potential applications. We leave these directions as future works.

7. Conclusion

We presented an efficient approach for customized video generation with textual and structural guidance. By employing a pre-trained image LDM as frozen spatial modules, our video generator exhibits a significant advantage in inheriting the wealth of visual concepts while maintaining satisfactory temporal coherence. Furthermore, we introduced a temporal masking mechanism to facilitate the synthesis of longer videos, which is a non-trivial extension for diffusion models. Ablation studies confirm the effectiveness of our proposed designs. Comparisons with state-of-the-art methods reveal our superiority in controllable text-to-video generation both quantitatively and qualitatively. Several applications show practical usage.

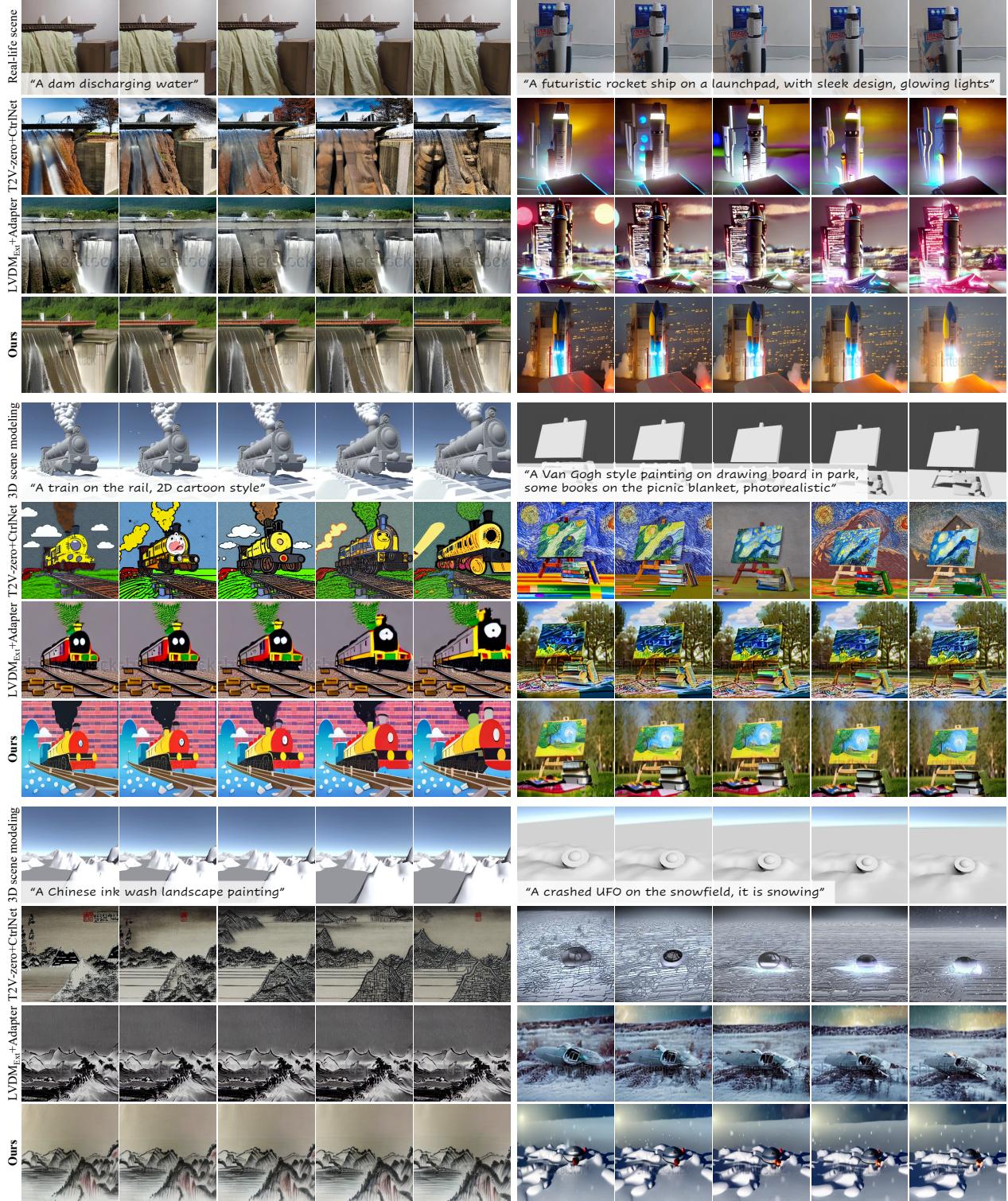


Figure 7. Visual comparison on videos generated in two applications, *i.e.*, real-life scene to video (top) and 3D scene modeling to video (middle & bottom).

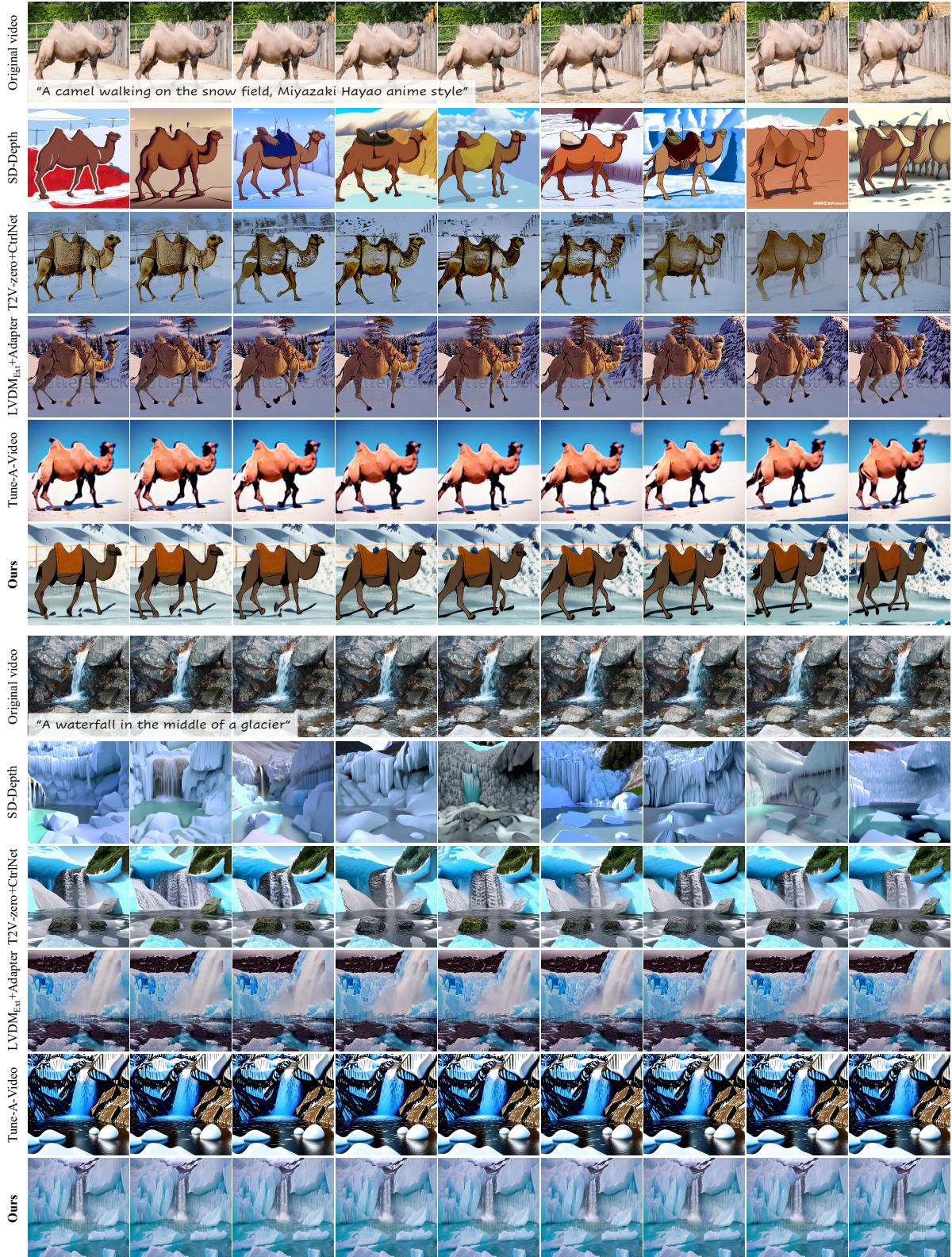


Figure 8. Visual comparison on examples of video re-rendering application, *i.e.*, ‘camel’ (top) and ‘waterfall’ (bottom).

References

- [1] Max Bain, Arsha Nagrani, G  l Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6
- [2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2
- [3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 8
- [4] 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. *arXiv preprint arXiv:2304.08818*, 2023. 3, 6
- [5] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3
- [6] Patrick Esser, Johnathan Chiu, Parmida Atighchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 3, 8
- [7] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [8] Xianfan Gu, Chuan Wen, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897*, 2023. 3
- [9] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 3, 6
- [10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations (ICLR)*, 2023. 3
- [11] 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
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2
- [13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Conference on Neural Information Processing Systems Workshop (NIPSW)*, 2021. 2, 6
- [14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video dif- fusion models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3
- [15] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations (ICLR)*, 2023. 6
- [16] Aapo Hyv  inen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 3
- [17] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 3
- [18] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3
- [19] Levon Khachatryan, Andranik Moossian, 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, 6
- [20] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Glichen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023. 2
- [21] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. 3
- [22] Siwei Lyu. Interpretation and generalization of score matching. *arXiv preprint arXiv:1205.2629*, 2012. 3
- [23] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 3
- [24] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 3
- [25] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3
- [26] Eyal Molad, Eliahu Horwitz, Dani Vavlevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 3
- [27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhonggang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 6
- [28] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion

- models. In *International Conference on Machine Learning (ICML)*, 2022. 2
- [29] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 8
- [30] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xiantao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023. 3
- [31] 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 (ICML)*, 2021. 2
- [32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 2
- [33] 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
- [34] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(3):1623–1637, 2020. 5
- [35] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020. 6
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 6
- [37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015. 4
- [38] 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. *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3
- [39] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *Conference on Neural Information Processing Systems Workshop (NIPSW)*, 2021. 4
- [40] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *ACL Conference of the North American Chapter of the Association for Computational Linguistics*, 2018. 5
- [41] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 6
- [42] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015. 2
- [43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 3, 6, 8
- [44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3
- [45] 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. 6
- [46] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3
- [47] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *International Conference on Learning Representations Workshop (ICLRW)*, 2019. 6
- [48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 5
- [49] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022. 3, 8
- [50] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6
- [51] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. *arXiv preprint arXiv:2211.15518*, 2022. 2
- [52] Yu Zeng, Zhe Lin, Jianming Zhang, Qing Liu, John Colomosse, Jason Kuen, and Vishal M Patel. Scenecomposer: Any-level semantic image synthesis. *arXiv preprint arXiv:2211.11742*, 2022. 2
- [53] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 6

- [54] 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. 3, 6