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

Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, *Senior Member, IEEE*, and Tien-Tsin Wong, *Member, IEEE*

***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 introduction 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.**

***Index Terms*—Content synthesis, diffusion models, temporal coherence, text-to-video generation.**

I. 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 [1], [2] 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 (Fig. 1 (top)). With 3D modeling software that supports depth map exportation, the physical objects can be replaced with simple geometric elements or any accessible 3D assets (Fig. 1 (middle)). Naturally, another alternative is to make use of the estimated depth from existing videos (Fig. 1 (bottom)). So, the combination of textual and structural guidance provides users with both flexibility and controllability to customize their videos.

To achieve this, we formulate the conditional video generation using a Latent Diffusion Model (LDM) [3] that adopts a diffusion model in a compressed lower-dimensional latent space. Training an open-world video generation model with text conditions presents two primary challenges: the necessity for considerable computational resources, which may be hard to accommodate, and the demand for text-video data encompassing a wealth of concepts. To tackle these, we propose separating the training of spatial modules (for image synthesis) and temporal modules (for temporal coherence). This design offers two benefits: (i) training the model components separately eases the computational resource requirements, which is especially crucial for resource-intensive tasks; (ii) as image datasets

Manuscript received 15 August 2023; revised 30 January 2024; accepted 6 February 2024. Date of publication 14 February 2024; date of current version 3 January 2025. Recommended for acceptance by T. Lee. *(Corresponding author: Menghan Xia.)*

Jinbo Xing, Yuxin Liu, Yuechen Zhang, Hanyuan Liu, and Tien-Tsin Wong are with The Chinese University of Hong Kong, Hong Kong (e-mail: jbxing@cse.cuhk.edu.hk; yxliu22@cse.cuhk.edu.hk; zhangyc@link.cuhk.edu.hk; liuhy@cse.cuhk.edu.hk; ttwong@cse.cuhk.edu.hk).

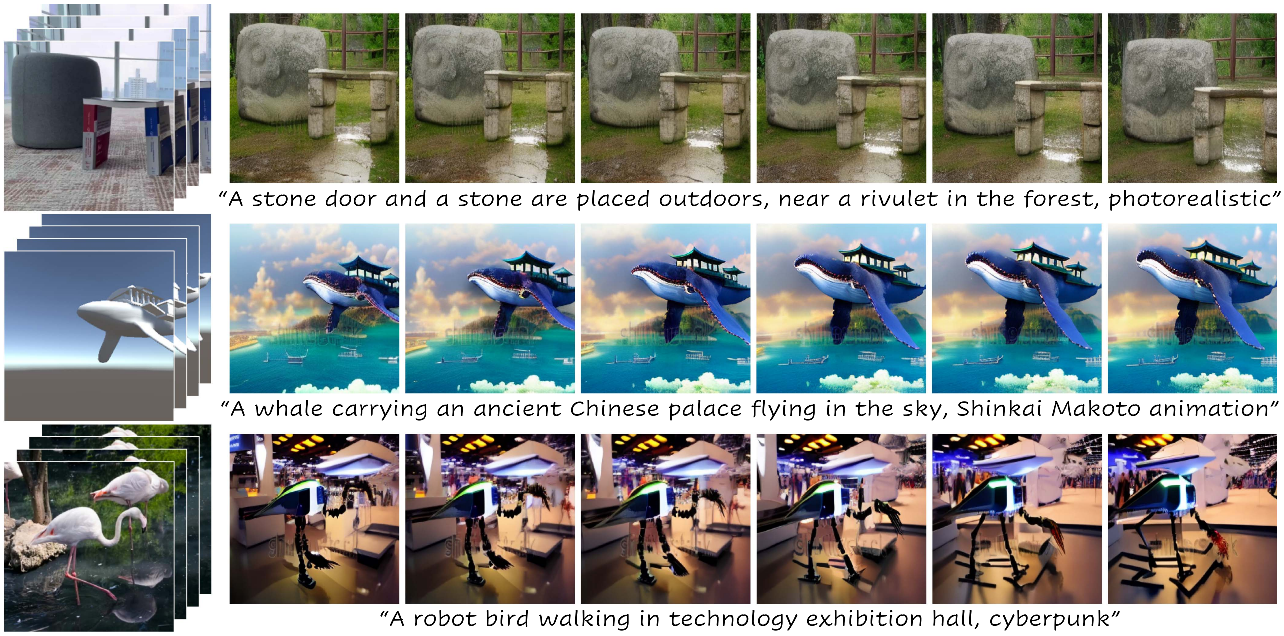
Menghan Xia, Yong Zhang, Haoxin Chen, Xiaodong Cun, Xintao Wang, and Ying Shan are with Tencent AI Lab., Shenzhen 518057, China (e-mail: menghanxyz@gmail.com; zhangyong201303@gmail.com; jszxchx@gmail.com; vinthony@gmail.com; xintao.alpha@gmail.com; yingsshan@tencent.com).

Yingqing He is with The Hong Kong University of Science and Technology, Hong Kong (e-mail: yhebm@connect.ust.hk).

The code, model weights, and videos are publicly available at our project page: https://doubiiu.github.io/projects/Make-Your-Video/

This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2024.3365804, provided by the authors.

Digital Object Identifier 10.1109/TVCG.2024.3365804



"A stone door and a stone are placed outdoors, near a rivulet in the forest, photorealistic"

"A whale carrying an ancient Chinese palace flying in the sky, Shinkai Makoto animation"

"A robot bird walking in technology exhibition hall, cyberpunk"

Fig. 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.

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. The main challenges lie in preserving the rich knowledge in the pre-trained image model and adapting it to achieve temporal coherence for video generation. 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× 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 effective 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.

II. Related Work

A. Diffusion Models for Text-to-Image Generation

Diffusion models [4], [5] (DMs) have recently shown unprecedented generative semantic and compositional power, attracting attention from both academia and industry. By absorbing the text embedding (e.g., from CLIP [6]), they have been successfully adopted for text-to-image generation (T2I) [3], [7], [8], [9]. GLIDE [9] introduces classifier-free guidance [10] to the text-conditioned image generation DMs, improving image quality in both photorealism and text-image alignment, which are further boosted by using CLIP [6] feature space in DALL·E 2. Moreover, a line of works [11], [12], [13] 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 address this, cascaded (Imagen [8]) and latent diffusion models (LDMs [3]) 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 cost, our method builds upon LDMs by introducing video awareness through additional learnable temporal modules and training on text-video data, while keeping the original weights of LDMs frozen.

B. Diffusion Models for Text-to-Video Generation

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 [1] proposes effective cascaded diffusion models and $\mathbf{v}$-prediction.

To reduce the training cost, many subsequent studies [2], [15], [16], [17] transfer T2I knowledge to T2V generation by initiating from pre-trained T2I models [18] and fine-tuning the entire model. Differently, T2v-zero [19] is proposed as 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 [20] and FollowYourPose [21] 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. Most recently, concurrent works ControlVideo [22] and VideoComposer [23] propose depth-conditioned T2V methods built upon a pre-trained T2I model, while they suffer from severe text-video misalignment issue due to lack of design for inheriting concepts in image model. As a concurrent work, Video-LDM [24] 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 [2], [24] and prediction [24], [25] 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 [24] and indoor scene [25]. Our work aims to efficiently and effectively adapt a pre-trained T2I model to a joint text-structure-guided video generator, and investigate general *diffusion-based video prediction mechanisms* for longer video synthesis.

C. Text-Driven Video Editing

In recent studies, DMs have demonstrated their efficacy in image editing tasks, as evidenced by several works [26], [27]. However, their application to video editing on individual frames can result in temporal inconsistency issues. To address this, Text2LIVE [28] and Lee et al. [29] combine Layered Neural Atlases [30] and the proposed text-driven image editing method, allowing texture-based video editing but struggling to reflect the intended edits accurately. To improve the video quality, recent diffusion-based video editing methods rely on either pre-trained large-scale video diffusion models [31], which are usually inaccessible and hard to reproduce due to the unaffordable training, or the inversion [32], [33], [34] followed by attention manipulation mechanism [35], [36], [37] using pre-trained T2I model, rendering tricky prompt engineering or manual hyper-parameter tuning process. Most recently, a line of diffusion-model-based works [38], [39] utilizing feature-level cross-frame correspondence achieves appealing editing results. 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 or leveraging RGB clues), whereas our model does not necessitate the source video as input.

III. 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.

A. 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 [40], [41], [42]. 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 $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, parameterized with learnable parameters θ; $t = 1 \ldots T$, 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 denoising objective is

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

where $\mathbf{c}$ is optional conditioning information (e.g., text prompt), t denotes a timestep uniformly sampled from $\{1, \ldots, T\}$, and $\boldsymbol{\epsilon}$ is the noise tensor used during the diffusion from $\mathbf{x}_0$ to $\mathbf{x}_t$. The neural backbone implementation of $\boldsymbol{\epsilon}_\theta(\circ; \mathbf{c}, t)$ is generally a 2D U-Net [43] with cross-attention conditioning mechanisms.

Latent diffusion models (LDMs): [3] 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 $\hat{x} = \mathcal{D}(\mathcal{E}(\mathbf{x})) \approx \mathbf{x}$. A DM can then be trained in the compressed latent space and turned into a LDM. The corresponding objective is similar to (1), except for replacing $\mathbf{x}$ with its latent representaton $\mathbf{z} = \mathcal{E}(\mathbf{x})$.

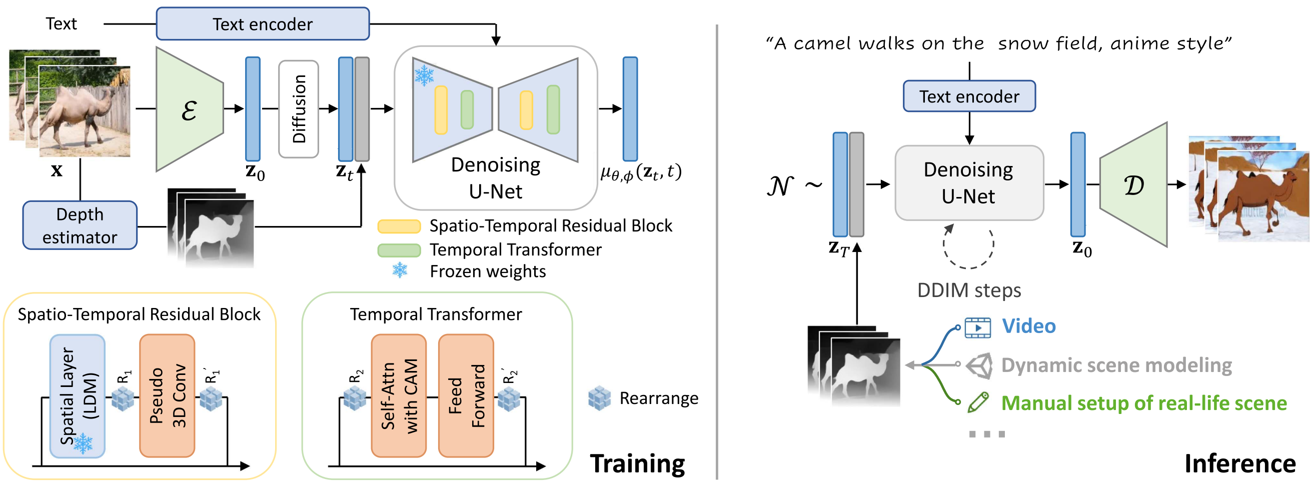


Fig. 2. Flowchart of the proposed method. During training (left), the input video $\mathbf{x}$ is first encoded into latent feature $\mathbf{z}_0$ with a fixed pre-trained encoder $\mathcal{E}$ and diffused to $\mathbf{z}_t$. Meanwhile, the depth sequence will be extracted with the off-the-shelf depth estimator MiDas and concatenated with $\mathbf{z}_t$, and the text is encoded by a frozen CLIP 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.

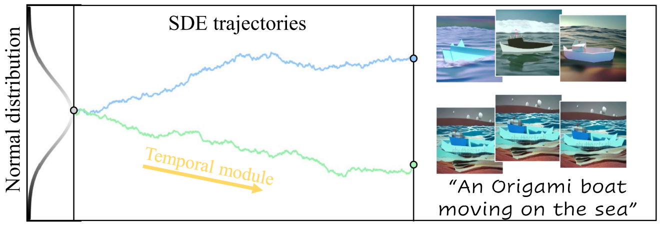


Fig. 3. Concept depiction of the effect of temporal modules $\boldsymbol{\epsilon}_\phi$. Given a batch of noises, $\boldsymbol{\epsilon}_\phi$ pushes the SDE trajectories toward target data distributions with cross-frame coherence. We draw a single trajectory for a batch for conciseness.

B. Adapting LDMs for Conditional T2V Generation

We employ an image 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 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 [44]. 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 Fig. 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 Fig. 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 L is the number of frames, 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 [45] with $\mathbf{z}_t$ that is diffused from latent feature $\mathbf{z}_0$. The textural guidance is introduced through cross-attention layers in the CLIP text embedding space.

We implement two temporal modules $\boldsymbol{\epsilon}_\phi$ to equip the pre-trained image CLDM with the ability to model temporal sequences, namely, Spatio-Temporal Residual Block and Temporal Transformer, shown at the bottom of Fig. 2.

Spatio-Temporal Residual Block (STRB): is an extension of the original 2D residual block containing spatial layers only. To make pre-trained image models capable of capturing temporal priors in videos, prior and concurrent works [2], [16], [24] typically incorporate only 1D temporal modeling layers (e.g., convolution or self-attention). However, these designs may be sub-optimal: (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, thus this domain adaptation could overwhelm light-weight 1D temporal modules; 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. Based on these considerations, we introduce additional learnable spatial convolution, making pseudo 3D conv, i.e., 2D spatial conv followed by 1D temporal conv, and the advantages in adapting image model to video generation model is evidenced in Section IV-C. In practical, 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. While STRB will regard the batched features as video data, and the corresponding rearranging operations in Fig. 2 are (using `einops` [46] notation):

$$\mathrm{R}_1: \texttt{rearrange}(\mathbf{z},\ \texttt{(b l) c h w} \rightarrow \texttt{b c l h w})$$

$$\mathrm{R}_1': \texttt{rearrange}(\mathbf{z},\ \texttt{b c l h w} \rightarrow \texttt{(b l) c h w}).$$

Temporal Transformer (TT): is located behind the original spatial transformer and is designed to exploit the property of temporal self-similarity to learn the inherent motion priors in video data. Concretely, it consists of temporal self-attention modules with learnable relative positional embeddings [47], making the network aware of temporal locations of frames, and a feed-forward layer [48]. TT only applies along the temporal axis, which is realized by the rearrangement of features (see the bottom of Fig. 2):

$$\mathrm{R}_2: \texttt{rearrange}(\mathbf{z},\ \texttt{(b l) c h w} \rightarrow \texttt{(b h w) l c})$$

$$\mathrm{R}_2': \texttt{rearrange}(\mathbf{z},\ \texttt{(b h w) l c} \rightarrow \texttt{(b l) c h w}).$$

The STRB and TT are initialized with an identify function, making the whole video model in the initial state work exactly the same as the original image model. Then our controllable T2V framework with denoiser $\boldsymbol{\epsilon}_{\theta,\phi}$ can be trained with a similar setting to the underlying CLDM. It is worth noting that the original weights $\boldsymbol{\epsilon}_{\theta}$ are frozen and only the added temporal modules $\boldsymbol{\epsilon}_{\phi}$ are learned via the optimization objective:

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

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

During inference, as depicted in Fig. 2 (right), the depth s obtained from multiple sources, e.g., estimated from existing videos or exported from 3 d scene modeling software, 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 a randomly sampled noise $\mathbf{z}_T \in \mathbb{R}^{L\times C\times H'\times W'}$ from Gaussian distribution and finally, the denoised latent $\mathbf{z}_0$ is converted into a video in pixel space using the pre-trained decoder $\mathcal{D}$.

C. Temporal Masking for Longer Video Synthesis

Our conditional video LDM can generate satisfactory videos with the same number of frames (i.e., 16 frames) as in the training phase. However, generating longer videos is highly valuable, but this aspect is under-exposed in *video diffusion models.* When using our model 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.

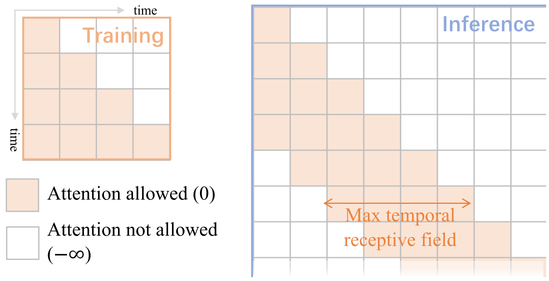


Fig. 4. Illustration of the causal attention mask during training & inference.

To alleviate this problem, we propose introducing 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 Fig. 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:

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

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 ($\mathbf{M}_{i,j} = 0$ 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 itself and the previous $L_M - 1$ tokens, where L_M is the maximum temporal receptive field and set to be the same as the training length (16 frames). 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.

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 IV-C. Thus we also adopt the causal attention mask in all experiments. Although be aware that the concept of causality is widely used in sequence modeling, e.g., video generation [49], [50], language modeling [51], and speech process [52], with distinct targets and domains, we adopt it to address the unique and inherent issue in extending *short video diffusion models* for *longer* video generation.

IV. Experiments

A. Implementation Details

Our development is based on depth-conditioned Latent Diffusion Models [3] (a.k.a Stable-Diffusion-Depth) implemented with PyTorch and the public pre-trained weights. We train the newly introduced layers with 50 K steps on the learning rate 1×10^{-4} and valid mini-batch size 512 with DeepSpeed [53]. At inference, we use DDIM sampler [33] with classifier-free guidance [10] in all experiments. More implementation details are provided in the *Supplement*.

TABLE I
QUANTITATIVE COMPARISONS (FVD AND KVD) 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	-
MagicVideo	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**

The best results are marked in bold.

Dataset: We use the WebVid-10M [54] dataset to turn the depth-conditioned LDM into a controllable text-to-video generator. WebVid-10 M consists of 10.7 million video-caption pairs with a total of 52 K 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.

B. 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 [16] and T2V-zero [19] are general text-to-video methods but capable of generating videos with additional conditions supported by ControlNet [12] or Adapter [13], 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 [17], MagicVideo [15], Make-A-Video [2], and Video LDM [24]. 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-to-video synthesis.

To evaluate the performance of video generation, we report the commonly-used Fréchet Video Distance (FVD) [55] and Kernel Video Distance (KVD) [55], 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 2 K model samples (16 frames) with samples from evaluation [15], [24] datasets, where we adopt commonly used UCF-101 [56] and MSR-VTT [57] for benchmarking. For UCF-101, we directly use UCF class names [24] as text conditioning, while for MSR-VTT, we randomly select one of the accompanied captions of each video from the dataset. We evaluate each error metric at the resolution of 256 × 256 with 16 frames.

We evaluate in the zero-shot setting and tabulate the quantitative performance on UCF-101 in Table I. 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 100 M text-video pairs, i.e., HD-VILA-100M [58]. 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 provided in the supplementary video and our project page[*], and also made in the context of several application scenarios, as shown in Figs. 7 and 8.

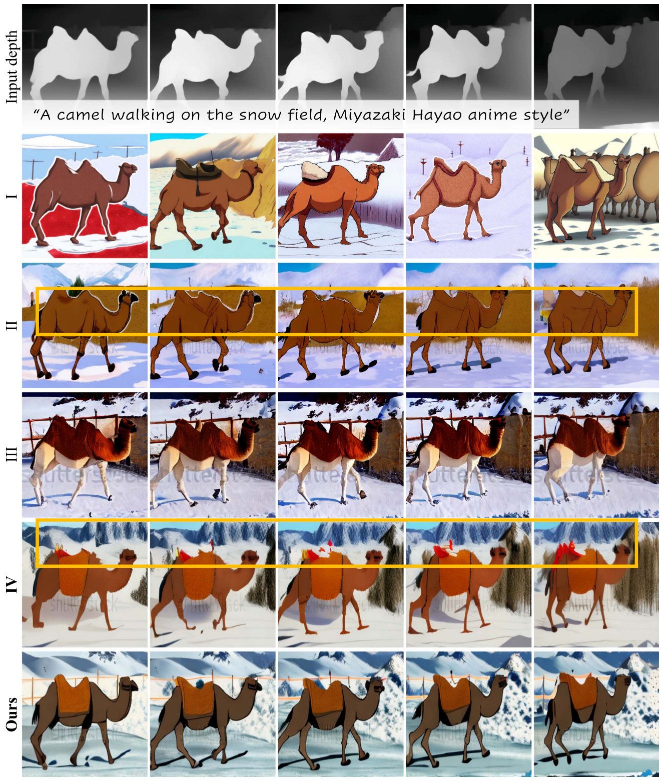

Input depth
"A camel walking on the snow field, Miyazaki Hayao anime style"
I
II
III
IV
Ours

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

In addition to FVD and KVD, we measure temporal consistency and prompt conformity on the depth-conditioned text-to-video methods for performance comparison. Following [20], 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. The results on UCF-101 are shown in Table II, where 'Temp.' and 'Prompt' indicate temporal coherence and prompt conformity, respectively. Our proposed method exhibits a substantial improvement over the previous approaches regarding both temporal coherence and text-video alignment, thanks to the effective design of temporal modeling modules and inheriting rich text-visual concepts from the pre-trained T2I model. In contrast, the limited cross-frame attention mechanism in T2V-zero+CtrlNet results in inferior temporal coherence, and full U-Net fine-tuning in LVDM_{Ext}+Adapter leads to prompt-visual

[*] https://doubiiu.github.io/projects/Make-Your-Video/

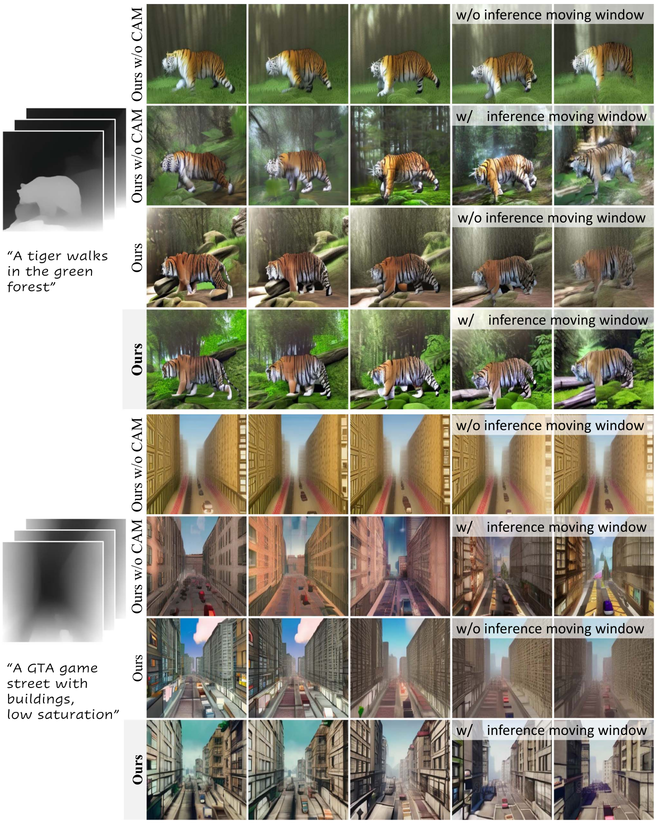

Ours w/o CAM
w/o inference moving window
Ours w/o CAM
w/ inference moving window
Ours
w/o inference moving window
Ours
w/ inference moving window
"A tiger walks in the green forest"
Ours w/o CAM
w/o inference moving window
Ours w/o CAM
w/ inference moving window
Ours
w/o inference moving window
Ours
w/ inference moving window
"A GTA game street with buildings, low saturation"

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

TABLE II
Quantitative Comparisons (Temporal Consistency and Prompt-Video Conformity) With State-of-the-Art Depth-Conditioned Text-to-Video Generation Methods on UCF-101 for the Zero-Shot Setting

Method	Condition	Temp. ↑	Prompt ↑
T2V-zero+CtrlNet	Text+Depth	0.8546	0.3018
LVDM$_{\text{Ext}}$+Adapter	Text+Depth	0.9444	0.2913
Ours w/o CAM	Text+Depth	0.9484	0.3063
Ours	Text+Depth	**0.9567**	**0.3107**

The best results are marked in bold.

TABLE III
Quantitative Comparisons With State-of-the-Art Methods on MSR-VTT for the Zero-Shot Setting

Method	FVD ↓	KVD ↓	Temp. ↑	Prompt ↑
T2V-zero+CtrlNet	677.35	69.18	0.8148	0.2816
LVDM$_{\text{Ext}}$+Adapter	492.53	65.44	0.9317	0.2715
Ours	**254.26**	**18.37**	**0.9398**	**0.2926**

The best results are marked in bold.

concept-forgetting issue, rendering low prompt-video conformity. In addition to UCF-101, the quantitative comparisons on MSR-VTT shown in Table III demonstrate consistent superiority of our model against the existing methods.

TABLE IV
Ablation Study of Adapting Strategies, Including Different Modules and Fine-Tuned Parameters, on UCF-101 for Zero-Shot Setting

Variant	Ft. para.	FVD ↓	KVD ↓	Temp. ↑	Prompt ↑
I. SD-Depth	None	1422.30	316.25	0.7238	**0.3365**
II. w/ TT	TT	803.24	89.26	0.8956	0.3068
III. w/ TT	Full U-Net	500.96	77.72	0.9529	0.2796
IV. w/ TT+TC	TT+TC	443.63	40.86	0.9397	0.2982
V. w/ TT+P3D (Ours)	TT+P3D	**330.49**	**29.52**	**0.9567**	0.3107

(Note: TT=temporal transformer, TC=1D temporal conv, P3D=pseudo 3D conv).
The best results are marked in bold.

C. 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 IV. By comparing baselines (I.), (II.), and (IV.), we can observe the improved performance in terms of FVD, KVD and temporal coherence by introducing more temporal modules that increase the temporal modeling capability. However, the prompt-video conformity decreases due to the domain shift and concept-forgetting. The comparison between (V.) and (IV.) highlights the benefits of incorporating additional spatial layers with TC, which enhances adaptability between the newly introduced temporal modules and the fixed spatial modules, rendering better temporal coherence and significantly improved prompt-video adherence. 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 the decreased prompt-video conformity and visually, the Fig. 5 where (III.) fails to reflect the 'anime style'. The figure can also qualitatively tells the superiority of our full method in terms of both temporal coherence and adherence 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 I (bottom), CAM can boost the performance of conditional video generation in a 16-frame setting, owing to the introduced directional video motion prior.

For longer video synthesis (64 frames), we directly apply the trained 'Ours w/o CAM' as a baseline, and besides, similar to the proposed solution in concurrent works of extending short video diffusion models for longer video generation [24], [59], we construct another baseline by conducting inference using

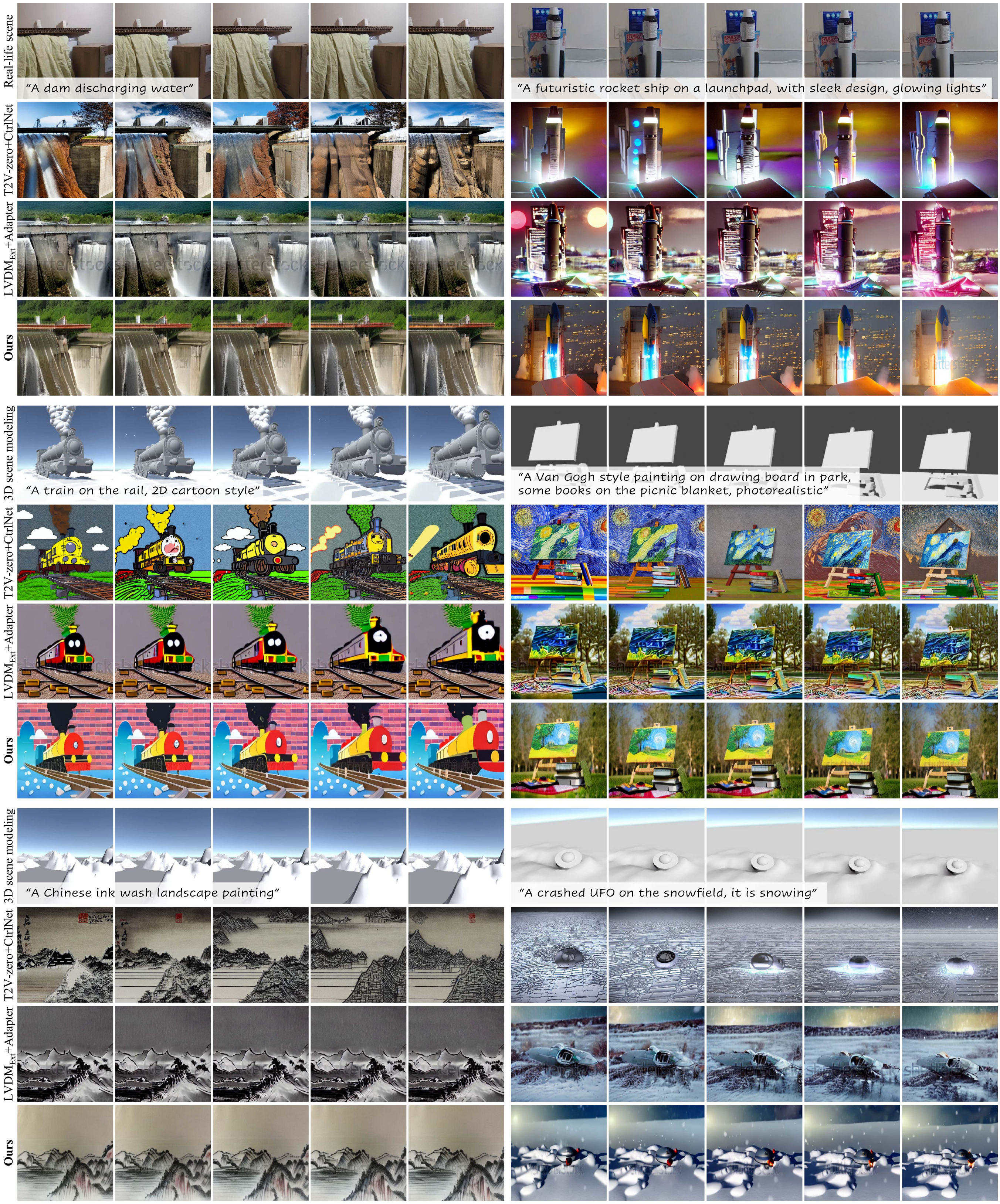

Real-life scene
"A dam discharging water"
"A futuristic rocket ship on a launchpad, with sleek design, glowing lights"
T2V-zero+CtrlNet
LVDM$_{Ext}$+Adapter
Ours
3D scene modeling
"A train on the rail, 2D cartoon style"
"A Van Gogh style painting on drawing board in park, some books on the picnic blanket, photorealistic"
T2V-zero+CtrlNet
LVDM$_{Ext}$+Adapter
Ours
3D scene modeling
"A Chinese ink wash landscape painting"
"A crashed UFO on the snowfield, it is snowing"
T2V-zero+CtrlNet
LVDM$_{Ext}$+Adapter
Ours

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

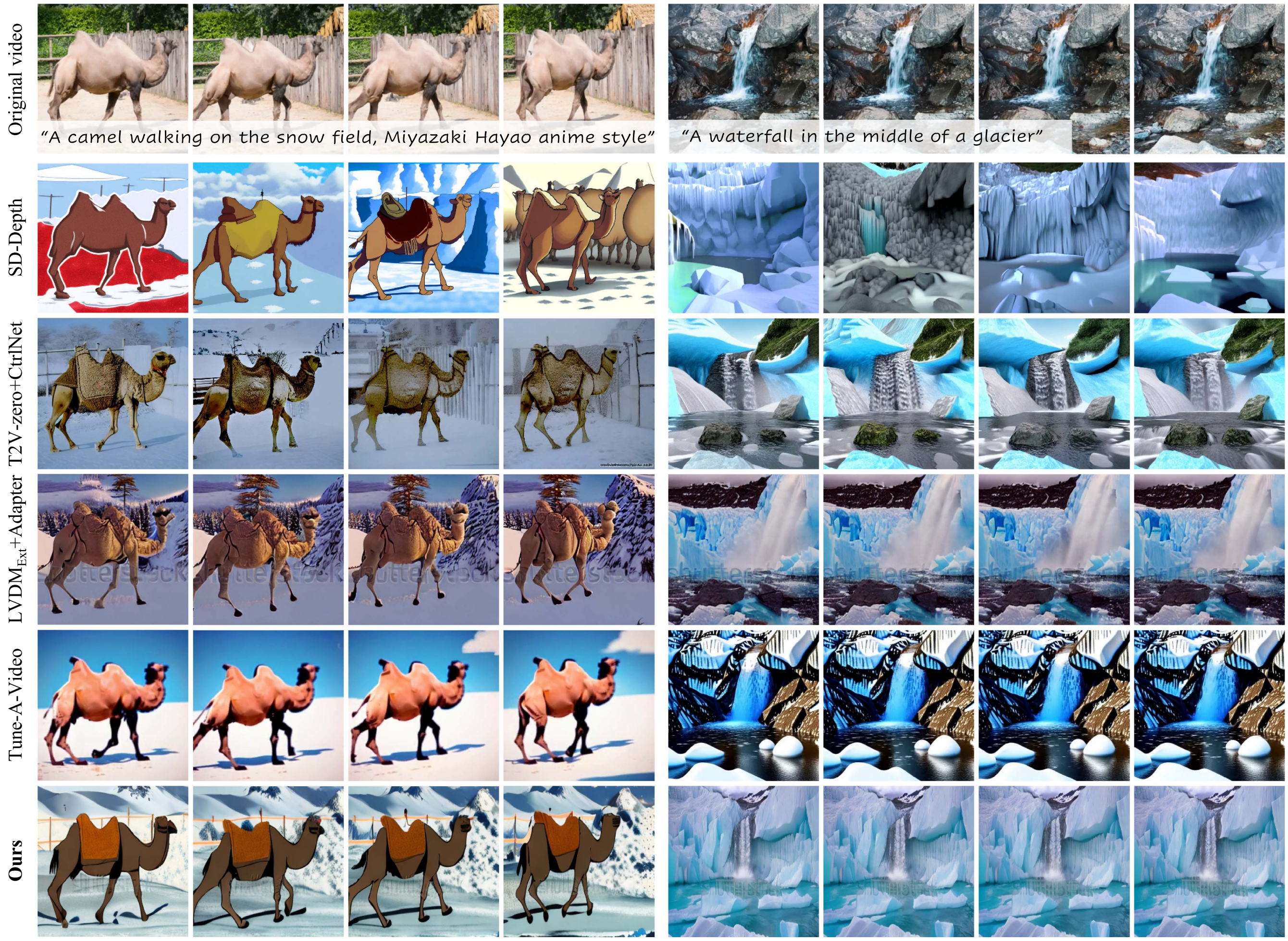

Original video
"A camel walking on the snow field, Miyazaki Hayao anime style"
"A waterfall in the middle of a glacier"
SD-Depth
T2V-zero+CtrlNet
LVDM$_{Ext}$+Adapter
Tune-A-Video
Ours

Fig. 8. Visual comparison on examples of video re-rendering application, i.e., "camel" (left) and 'waterfall' (right).

TABLE V
Ablation Study of Causal Attention Mask in Generating Longer Videos (64 Frames) on UCF-101 for Zero-Shot Setting

Model	Infer. window	FVD ↓	KVD ↓	Temp. ↑	Prompt ↑
Ours w/o CAM	✗	990.71	98.40	0.9159	0.2873
	✓	569.93	69.58	0.9366	0.3064
Ours	✗	486.88	48.78	0.9400	0.3086
	✓	**436.70**	**44.45**	**0.9448**	**0.3101**

FVD and KVD are the averaged results computed on non-overlapping 16-frame clips. (Note: 'infer. window' indicates the moving window inference strategy.).
The best results are marked in bold.

'Ours w/o CAM' model in a moving window (16 frames) style. In this manner, the frame length at inference and training can be maintained. In addition, we construct a baseline 'Ours w/o Infer. window' by removing moving window inference strategy from our final method. The quantitative evaluation result in Table V shows that the moving window inference strategy can boost the performance of 'Ours w/o CAM' model by roughly aligning the training and inference setting. However, two three variants are inferior compared with our full model in all metrics. Moreover, removing moving window from our final method (the 3 rd row in Table V) leads to reduced performance as well.

The qualitative comparison is presented in Fig. 6, wherein our baseline variant 'Ours w/o CAM' without moving window trick suffers from severe quality degradation and temporal inconsistency due to the mismatched training and inference length. While adopting the moving window strategy during inference could mitigate its quality degradation issue, it still fails to generate coherent frames in the global context due to the lack of causality among windows (more evident in the supplementary video). Moreover, 'Ours' without moving window suffers from severe quality degradation in those subsequent frames after the 16th frame (note that the training video length is 16). In contrast, our method could produce plausible longer videos with greater detail and improved cross-frame coherence.

V. Applications

A. 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 Figs. 1 (top) and 7 (top). In comparison to other

TABLE VI
QUANTITATIVE COMPARISONS FOR VIDEO RE-RENDERING

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

The best results are marked in bold and the second-best results are marked in underline.

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 dynamic scenes with 3D modeling software, e.g., Unity, and then *export* the motion structure (depth) for customized video generation, as shown in Figs. 1 (middle) and 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 Fig. 7. It is worth noting that our method can achieve both decent text alignment and cross-frame coherence.

B. 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) that is used by our model as spatial modules, and a video editing method, Tune-A-Video [35] combined with DDIM inversion [33], which requires per-video optimization and the original RGB video for inversion operation. Following previous works [20], [28], [35], we use videos from DAVIS [60] and other in-the-wild videos, from which 11 representative videos with manually designed text prompts are utilized for evaluation.

Quantitative evaluation: Table VI shows the results of each model. 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 Fig. 8.

Qualitative evaluation: We visually compare our method with other competitors in Fig. 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 project page for better comparison.

TABLE VII
USER STUDY STATISTICS OF *AVERAGE RANKING*↓ AND PREFERENCE RATE↑

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%**

The best results are marked in bold.

TABLE VIII
QUANTITATIVE COMPARISONS WITH CONCURRENT WORKS AND GEN-1 FOR VIDEO RE-RENDERING

Metric	ControlVideo	VideoComposer	Gen-1	Ours
Temp.↑	0.9803	0.9721	**0.9809**	0.9698
Prompt↑	0.3318	0.3502	0.3496	**0.3547**

The best results are marked in bold.

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, 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 VII. Our method earns the best ranking scores and preference rates in all three aspects.

VI. COMPARISON WITH CONCURRENT WORKS AND GEN-1

We also compare our proposed method with concurrent works on video re-rendering, which are two depth-conditioned text-to-video generation approaches (i.e., ControlVideo [22] and VideoComposer [23]). Additionally, we include Gen-1 [20] as a comparative method, which is a proprietary commercial web application with a far-reaching impact on video editing.

The evaluation setting is the same as the one described in Section V-B, and the comparison result is presented in Table VIII. Our method is on par with VideoComposer in temporal consistency, while being inferior to ControlVideo and Gen-1 in quantitative evaluations. However, we observe that ControlVideo tends to generate over-smoothed backgrounds with a severe cross-frame copy-and-paste effect, leading to a tricky quantitative temporal consistency performance, which can be seen in Fig. 9 (2nd row). In addition, its temporal inconsistency can be revealed from the visual results, e.g., the trees in the background of 'camel' case and the pattern of the cow head in the 'cow' case. In contrast, our method is capable of producing perceptually more coherent frames. Gen-1, as a commercial product, can generate

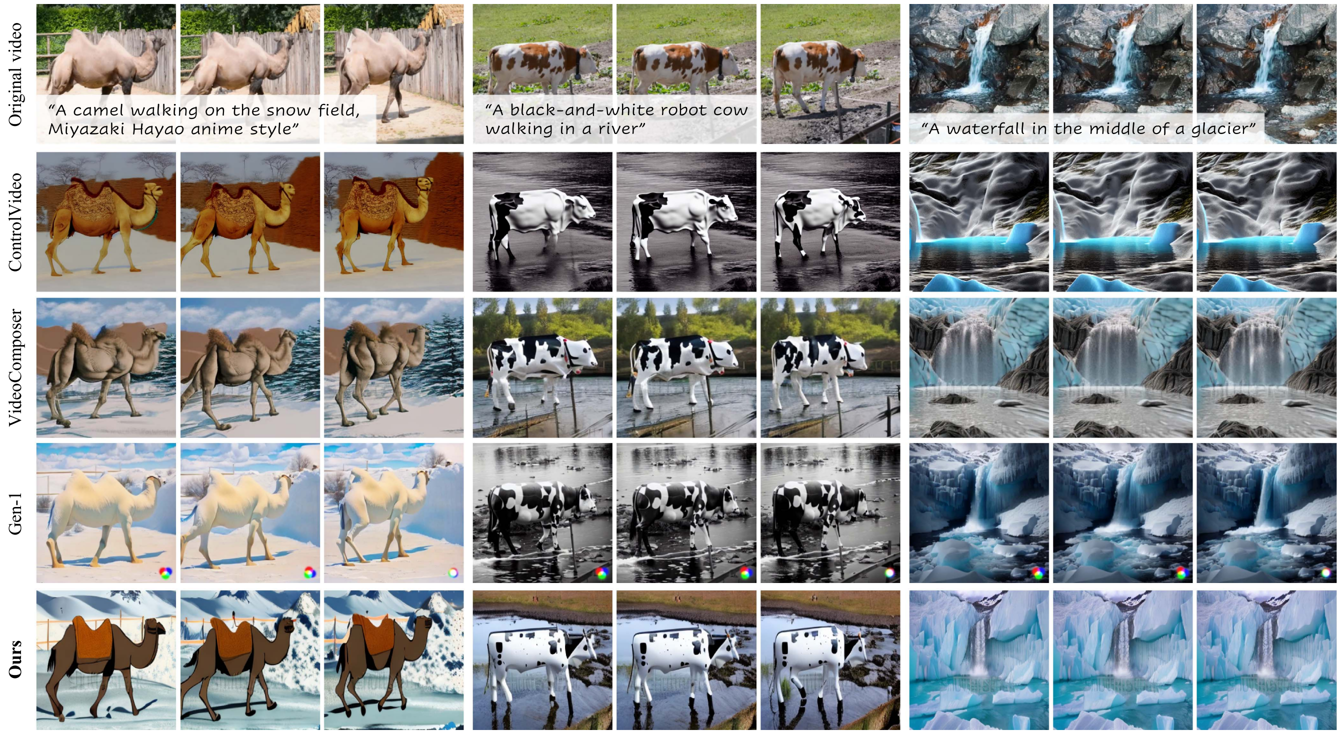

Original video
ControlVideo
VideoComposer
Gen-1
Ours
"A camel walking on the snow field, Miyazaki Hayao anime style"
"A black-and-white robot cow walking in a river"
"A waterfall in the middle of a glacier"

Fig. 9. Visual comparison with concurrent works and Gen-1 on examples of video re-rendering application, i.e., "camel" (left), "cow" (middle), and "waterfall" (right).

videos with good frame quality and temporal coherence. However, we point out that the statistics of online Gen-1 above should be very different from their arXiv version [20], as we found its current online version no longer supports 8-frame 256x448 video input (its original paper's setting) and may contain other unpublished mechanisms/techniques. We emphasize that Gen-1 utilizes much more computing resources (at least 115 k training steps with video-batchsize=1152 and image-batchsize=9216 on proprietary datasets in the paper).

Essentially, our method significantly outperforms other approaches in terms of prompt-video conformity, as shown in Table VIII and further visually evidenced in Fig. 9, e.g., in the 'camel' case, all other baselines fail to reflect the 'anime style'; and in the 'cow' case, neither of them can show the 'robot' concept, and ControlVideo & Gen-1 even wrongly synthesize monochrome videos due to the misunderstanding of the word 'black-and-white' describing cow pattern. Moreover, our method exhibits superior structure preservation in the 'waterfall' case compared to ControlVideo and VideoComposer. These advantages of our method demonstrate the success of the proposed adapting strategies in better inheriting the rich concepts and structure preservation from pre-trained image diffusion models.

VII. Discussion

A. Diversity of Generated Videos

Our proposed model accommodates two types of control signals (i.e., text prompts and frame-wise depth maps), which play different roles during generation. Specifically, the text prompt describes the video appearance and depth sequence specifies the overall motion structure. We present the qualitative results generated by utilizing 'single motion structure + multiple texts' in Fig. 10 and 'single text + multiple motion structures' in Fig. 11. These results demonstrate the diversity of the generated videos.

B. Impact of Depth Maps on Generation

To further explore the impact of input depth maps on synthesis results, we conduct the following experiments by modifying the depth maps (Note that zero values in depth maps indicate infinitely distant regions, and vice versa.):

- all-zero, all-half, all-one: setting all input depth map values to 0, 0.5, and 1, respectively,
- Gaussian noise, Gaussian blur,
- spatially half-zero, temporally half-zero, temporally interleave-zero: setting input depth map values to 0 in the spatially upper-half, temporally latter-half, and temporally interleaving manner, respectively.

As demonstrated in Fig. 12, **all-zero** and **all-half** tend to produce background concept, i.e., 'forest', and struggle to reflect the 'tiger' concept, as 'forest' may frequently be associated with near-zero depth maps during training. In contrast, **all-one** results in a whole 'tiger' video in an extreme close-up view. Furthermore, adding **Gaussian noise** to the depth maps can significantly impair the model's performance, as shown in the 3 rd row of Fig. 12, where the model incorrectly associate

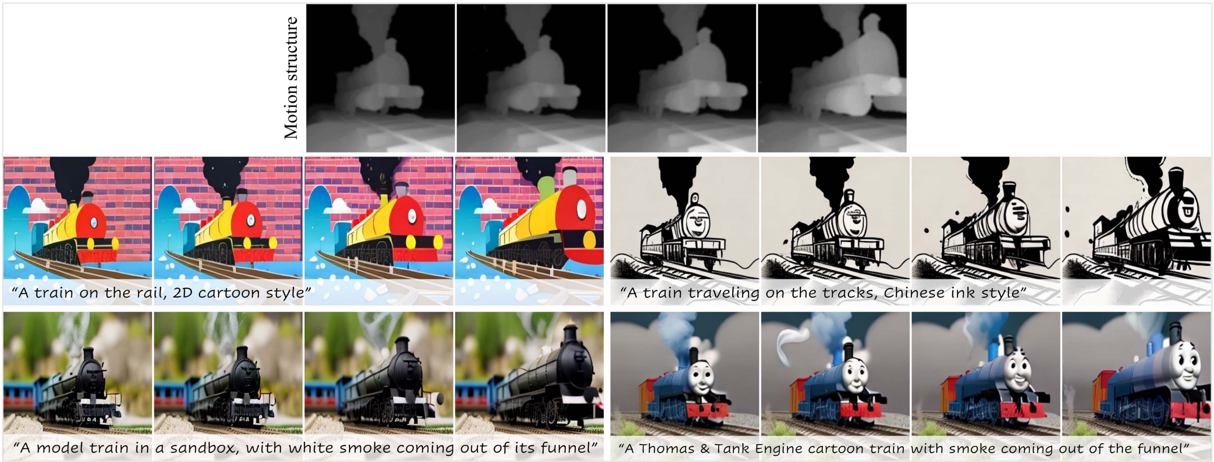

Motion structure
"A train on the rail, 2D cartoon style"
"A train traveling on the tracks, Chinese ink style"
"A model train in a sandbox, with white smoke coming out of its funnel"
"A Thomas & Tank Engine cartoon train with smoke coming out of the funnel"

Fig. 10. Visual results generated by our model using a single motion structure (top) and multiple different texts.

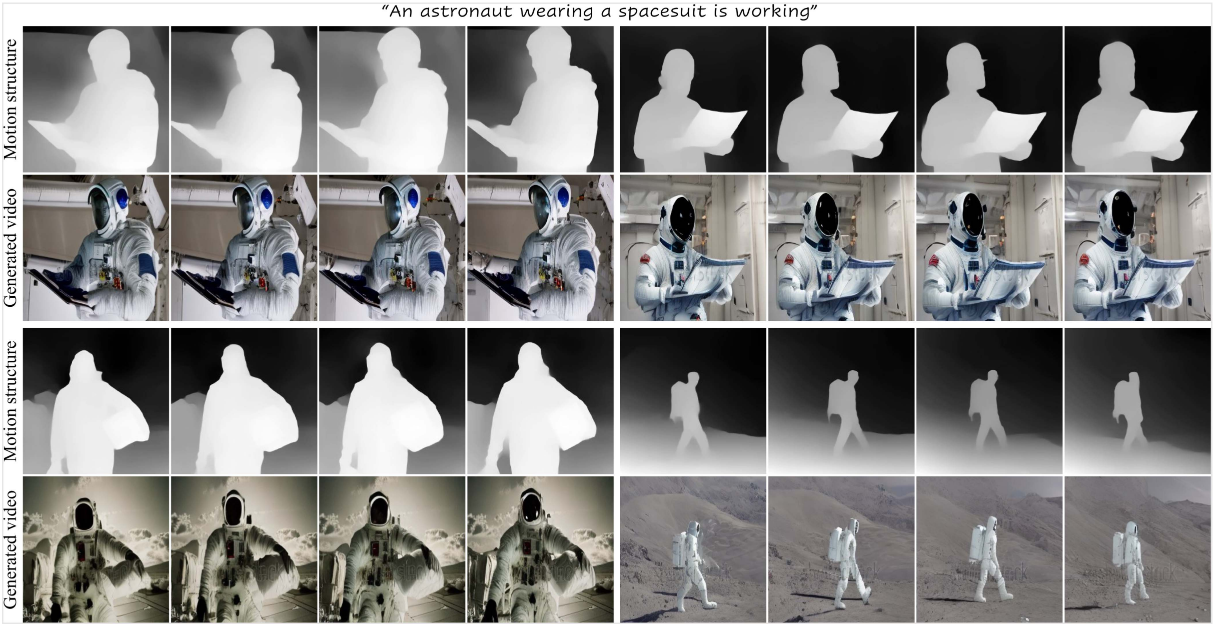


Fig. 11. Visual results generated by our model using a single text prompt (top) and multiple different motion structures.

different concepts with the depth maps, i.e., it wrongly produce trees in the bear motion structure region. Adding **Gaussian blur** may introduce artifacts around the object. While setting **spatially upper-half** region and **temporally latter-half** depth sequences to zero will produce background concepts in the text prompt like 'forest'. However, setting depth sequences to zero in a **temporally interleaving** manner does not significantly impact the generated videos while only introducing slight artifacts, as shown in the last row of Fig. 12. It is worth noting that all these modified input depth maps are out-of-distribution considering the distribution mapping during training, while our model still demonstrates a certain level of robustness.

C. Limitation

Our approach may experience temporal artifacts in the generated videos when the motion magnitude in the structural guidance is relatively large. We conjecture that this may be due to several factors: (i) The depth condition for training (and inference for the video-rerendering application) is estimated on a frame-by-frame basis and only provides 'relative' geometry information within a single frame, without taking into account the entire video content. This may introduce inconsistency guidance signals, hindering the model's ability to generate temporally consistent frames. The issue is exacerbated when motion magnitudes are larger; (ii) The temporal modeling capacity is limited by small receptive fields, especially when the inconsistent structural condition is present, rendering the failure to synthesize visual pleasing results for such cases. Additional specific limitations, including inability to synthesize videos featuring a specific individual, requirement for frame-wise depth guidance, and low-resolution results & flickering artifacts in high-frequency regions, can be found in the supplement.

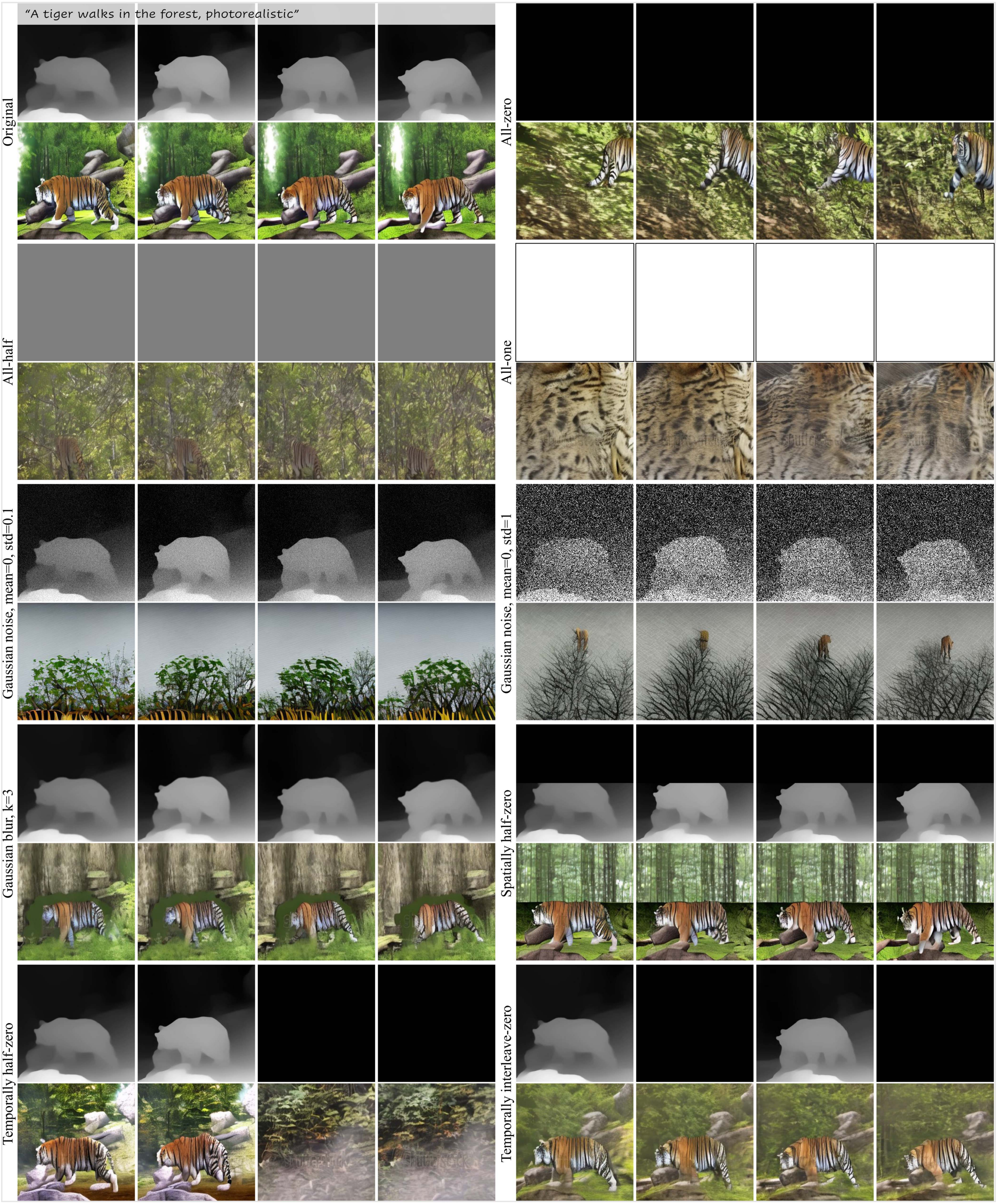

"A tiger walks in the forest, photorealistic"
Original
All-zero
All-half
All-one
Gaussian noise, mean=0, std=0.1
Gaussian noise, mean=0, std=1
Gaussian blur, k=3
Spatially half-zero
Temporally half-zero
Temporally interleave-zero

Fig. 12. Visual results of impact of modifying input depth-maps on the generated videos, i.e., original depth maps, all-zero, all-half, all-one (setting all input depth map values to 0, 0.5, and 1, respectively), Gaussian noises, Gaussian blur, spatially half-zero, temporally half-zero, and temporally interleave-zero. We show (modified) depth maps at the top and the corresponding generated videos at the bottom.

VIII. Conclusion

In this study, 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 temporal coherence. Additionally, we introduced a temporal masking mechanism to facilitate the synthesis of longer videos. Quantitative and qualitative comparisons with the state-of-the-arts demonstrate the superiority of our controllable text-to-video generation approach.

References

[1] J. Ho et al., "Imagen video: High definition video generation with diffusion models," 2022, *arXiv:2210.02303*.

[2] U. Singer et al., "Make-a-video: Text-to-video generation without text-video data," in *Proc. Int. Conf. Learn. Representations*, 2023.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10674–10685.

[4] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[6] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

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

[8] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Conf. Neural Inf. Process. Syst.*, 2022, pp. 36479–36494.

[9] A. Q. Nichol et al., "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 16784–16804.

[10] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Proc. Conf. Neural Inf. Process. Syst. Workshop*, 2021, pp. 217–232.

[11] Y. Li et al., "GLIGEN: Open-set grounded text-to-image generation," 2023, *arXiv:2301.07093*.

[12] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023, *arXiv:2302.05543*.

[13] C. Mou et al., "T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," 2023, *arXiv:2302.08453*.

[14] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Proc. Conf. Neural Inf. Process. Syst.*, 2022.

[15] D. Zhou, W. Wang, H. Yan, W. Lv, Y. Zhu, and J. Feng, "MagicVideo: Efficient video generation with latent diffusion models," 2022, *arXiv:2211.11018*.

[16] Y. He, T. Yang, Y. Zhang, Y. Shan, and Q. Chen, "Latent video diffusion models for high-fidelity video generation with arbitrary lengths," 2022, *arXiv:2211.13221*.

[17] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "CogVideo: Large-scale pretraining for text-to-video generation via transformers," in *Proc. Int. Conf. Learn. Representations*, 2023.

[18] Z. Luo et al., "VideoFusion: Decomposed diffusion models for high-quality video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10209–10218.

[19] L. Khachatryan et al., "Text2Video-zero: Text-to-image diffusion models are zero-shot video generators," 2023, *arXiv:2303.13439*.

[20] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," 2023, *arXiv:2302.03011*.

[21] Y. Ma et al., "Follow your pose: Pose-guided text-to-video generation using pose-free videos," 2023, *arXiv:2304.01186*.

[22] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, "ControlVideo: Training-free controllable text-to-video generation," 2023, *arXiv:2305.13077*.

[23] X. Wang et al., "VideoComposer: Compositional video synthesis with motion controllability," 2023, *arXiv:2306.02018*.

[24] A. Blattmann et al., "Align your latents: High-resolution video synthesis with latent diffusion models," 2023, *arXiv:2304.08818*.

[25] X. Gu, C. Wen, J. Song, and Y. Gao, "Seer: Language instructed video prediction with latent diffusion models," 2023, *arXiv:2303.14897*.

[26] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," in *Proc. Int. Conf. Learn. Representations*, 2023.

[27] B. Kawar et al., "Imagic: Text-based real image editing with diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6007–6017.

[28] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, "Text2Live: Text-driven layered image and video editing," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 707–723.

[29] Y.-C. Lee, J.-Z. G. Jang, Y.-T. Chen, E. Qiu, and J.-B. Huang, "Shape-aware text-driven layered video editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14317–14326.

[30] Y. Kasten, D. Ofri, O. Wang, and T. Dekel, "Layered neural atlases for consistent video editing," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–12, 2021.

[31] E. Molad et al., "Dreamix: Video diffusion models are general video editors," 2023, *arXiv:2302.01329*.

[32] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6038–6047.

[33] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Representations*, 2021.

[34] M. Zhao, R. Wang, F. Bao, C. Li, and J. Zhu, "ControlVideo: Adding conditional control for one shot text-to-video editing," 2023, *arXiv:2305.17098*.

[35] J. Z. Wu et al., "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," 2022, *arXiv:2212.11565*.

[36] C. Qi et al., "Fatezero: Fusing attentions for zero-shot text-based video editing," 2023, *arXiv:2303.09535*.

[37] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia, "Video-P2P: Video editing with cross-attention control," 2023, *arXiv:2303.04761*.

[38] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel, "Tokenflow: Consistent diffusion features for consistent video editing," 2023, *arXiv:2307.10373*.

[39] S. Yang, Y. Zhou, Z. Liu, and C. C. Loy, "Rerender a video: Zero-shot text-guided video-to-video translation," 2023, *arXiv:2306.07954*.

[40] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Representations*, 2021.

[41] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching," *J. Mach. Learn. Res.*, vol. 6, no. 4, 2005.

[42] S. Lyu, "Interpretation and generalization of score matching," 2012, *arXiv:1205.2629*.

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[44] C. Schuhmann et al., "Laion-400M: Open dataset of clip-filtered 400 million image-text pairs," in *Proc. Conf. Neural Inf. Process. Syst. Workshop*, 2021.

[45] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022.

[46] A. Rogozhnikov, "Einops: Clear and reliable tensor manipulations with einstein-like notation," in *Proc. Int. Conf. Learn. Representations*, 2022.

[47] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. ACL Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2018.

[48] A. Vaswani et al., "Attention is all you need," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017.

[49] S. Ge et al., "Long video generation with time-agnostic VQGAN and time-sensitive transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 102–118.

[50] R. Villegas et al., "Phenaki: Variable length video generation from open domain textual description," in *Proc. Int. Conf. Learn. Representations*, 2023.

[51] Z. Luo et al., "DecBERT: Enhancing the language understanding of BERT with causal attention masks," in *Proc. ACL Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2022.

[52] A. Nicolson and K. K. Paliwal, “Masked multi-head self-attention for causal speech enhancement,” *Speech Commun.*, vol. 125, no. 1, pp. 80–96, 2020.

[53] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, “DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters,” in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 3505–3506.

[54] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, “Frozen in time: A joint video and image encoder for end-to-end retrieval,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1708–1718.

[55] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, “FVD: A new metric for video generation,” in *Proc. Int. Conf. Learn. Representations Workshop*, 2019.

[56] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” 2012, *arXiv:1212.0402*.

[57] J. Xu, T. Mei, T. Yao, and Y. Rui, “MSR-VTT: A large video description dataset for bridging video and language,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5288–5296.

[58] H. Xue et al., “Advancing high-resolution video-language representation with large-scale video transcriptions,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5036–5045.

[59] F.-Y. Wang, W. Chen, G. Song, H.-J. Ye, Y. Liu, and H. Li, “Gen-L-Video: Multi-text to long video generation via temporal co-denoising,” 2023, *arXiv:2305.18264*.

[60] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, “The 2017 davis challenge on video object segmentation,” 2017, *arXiv: 1704.00675*.

Jinbo Xing received the BSc and MSc degrees from the Chinese University of Hong Kong, in 2020 and 2021 respectively. He is currently working toward the PhD degree in the Department of Computer Science and Engineering, the Chinese University of Hong Kong. His current research interests include 3D facial animation and AI-generated content (video generation).

Menghan Xia received the BEng degree in photogrammetry and remote sensing and the MEng degree in pattern recognition and intelligent system from Wuhan University, in 2014 and 2017, respectively, and the PhD degree in computer science from the Chinese University of Hong Kong. He is currently a senior researcher with Tencent AI Lab. His research interests include computer vision, image processing, and video generation.

Yuxin Liu received the BSc degree in computer science with the Chinese University of Hong Kong, in 2020. He is currently working toward the PhD degree in the Department of Computer Science and Engineering from the Chinese University of Hong Kong. His research interest includes computer graphics, stylized rendering, 3D content generation, and computer vision.

Yuechen Zhang received the BSc degree from the Chinese University of Hong Kong, in 2021. He is currently working toward the PhD degree in the Department of Computer Science and Engineering, the Chinese University of Hong Kong. His current research interests include controllable generation and stylization in computer vision.

Yong Zhang received the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2018. He is currently a senior researcher with Tencent AI Lab. From 2015 to 2017, he was a visiting scholar with the Rensselaer Polytechnic Institute. His research interests include computer vision and machine learning.

Yingqing He received the BSc degree from Beijing Forestry University, in 2018 and the MSc degree from the Hong Kong University of Science and Technology, in 2019. She is currently working toward the PhD degree with the Hong Kong University of Science and Technology. Her current research focuses on AI-generated content, particularly image/video synthesis and editing.

Hanyuan Liu received the BEng degree from Zhejiang University, in 2018. He is currently the PhD degree in the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His current research interests include computational photography, computational arts, computer graphics, generative models, and multimodal learning.

Haoxin Chen received the BEng degree, in 2019 and the MEng degree, in 2022, both in computer science and engineering from the South China University of Technology. He is currently a research engineer with Tencent AI Lab. His research interests include computer vision, video generation, and high-performance computing.

Xiaodong Cun received the BSc degree in computer science from Xidian University, and the MS and PhD degrees from the Department of Computer and Information Science, University of Macau, in 2018 and 2021, respectively. He is currently a senior researcher in the Visual Computing Center of Tencent AI Lab. His current research interests include image/video generation, translation, and editing.

Xintao Wang received the PhD degree from the Department of Information Engineering, the Chinese University of Hong Kong, in 2020. He is currently a senior researcher in Applied Research Center (ARC) with Tencent PCG and Tencent AI Lab. He was selected as an outstanding reviewer in CVPR 2019 and an outstanding reviewer (honorable mention) in BMVC 2019. He won the first place in several international super-resolution challenges, including NTIRE2019, NTIRE2018, and PIRM2018. His research interests include computer vision and deep learning.

Ying Shan (Senior Member, IEEE) is a distinguished scientist with Tencent, the director of the ARC Lab with Tencent PCG, and the director of the Visual Computing Center with Tencent AI Lab. Before joining Tencent, he worked with Microsoft Research as a post-doc researcher, SRI International (Sarnoff Subsidiary) as a Senior MTS, and Microsoft Bing Ads as a Principal Scientist Manager. He has published more than 100 papers in top conferences and journals in the areas of computer vision, machine learning, and data mining, served as ACs of CVPR and senior PC of KDD.

Tien-Tsin Wong (Member, IEEE) received the BSc, MPhil, and PhD degrees in computer science from the Chinese University of Hong Kong, in 1992, 1994, and 1998, respectively. He is currently a professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His main research interests include computer graphics, computational manga, precomputed lighting, image-based rendering, GPU techniques, medical visualization, multimedia compression, and computer vision. He received the *IEEE Transactions on Multimedia* Prize Paper Award 2005 and the Young Researcher Award 2004.