ControlVideo: Training-free Controllable Text-to-Video Generation

Yabo Zhang[1] **Yuxiang Wei**[1] **Dongsheng Jiang**[2] **Xiaopeng Zhang**[2] **Wangmeng Zuo**[1] (✉) **Qi Tian**[2]

[1]Harbin Institute of Technology [2]Huawei Cloud

Abstract

Text-driven diffusion models have unlocked unprecedented abilities in image generation, whereas their video counterpart still lags behind due to the excessive training cost of temporal modeling. Besides the training burden, the generated videos also suffer from appearance inconsistency and structural flickers, especially in long video synthesis. To address these challenges, we design a *training-free* framework called **ControlVideo** to enable natural and efficient text-to-video generation. ControlVideo, adapted from ControlNet, leverages coarsely structural consistency from input motion sequences, and introduces three modules to improve video generation. Firstly, to ensure appearance coherence between frames, ControlVideo adds fully cross-frame interaction in self-attention modules. Secondly, to mitigate the flicker effect, it introduces an interleaved-frame smoother that employs frame interpolation on alternated frames. Finally, to produce long videos efficiently, it utilizes a hierarchical sampler that separately synthesizes each short clip with holistic coherency. Empowered with these modules, ControlVideo outperforms the state-of-the-arts on extensive motion-prompt pairs quantitatively and qualitatively. Notably, thanks to the efficient designs, it generates both short and long videos within several minutes using one NVIDIA 2080Ti. Code is available at https://github.com/YBYBZhang/ControlVideo.

1 Introduction

Large-scale diffusion models have made a tremendous breakthrough on text-to-image synthesis [1, 22, 26, 29, 32] and their creative applications [6, 8, 21, 37]. Several works [5, 9, 11, 12, 34] attempt to replicate this success in the video counterpart, *i.e.*, modeling higher-dimensional complex video distributions in the wild world. However, training such a text-to-video model requires massive amounts of high-quality videos and computational resources, which limits the further research and applications by relevant communities.

To reduce the excessive training requirements, we study a new and efficient form: *controllable text-to-video generation with text-to-image models*. This task aims to produce a video conditioned on both a textual description and motion sequences (*e.g.*, depth or edge maps). As shown in Fig. 1, instead of learning the video distribution from scratch, it could efficiently leverage the generation capability of pre-trained text-to-image generative models [26, 29] and coarsely temporal consistency of motion sequences to produce vivid videos.

Recent studies [15, 40] have explored leveraging the structure controllability of ControlNet [43] or DDIM inversion [35] for video generation. Rather than synthesizing all frames independently, [15, 40] enhance appearance coherence by replacing original self-attention with the sparser cross-frame attention. Nevertheless, their video quality is still far behind photo-realistic videos in terms of: (i) inconsistent appearance between some frames (see Fig. 4 (a)), (ii) visible artifacts in large motion videos (see Fig. 4 (b)), *and* (iii) structural flickers during inter-frame transitions. For (i) and

Preprint. Under review.

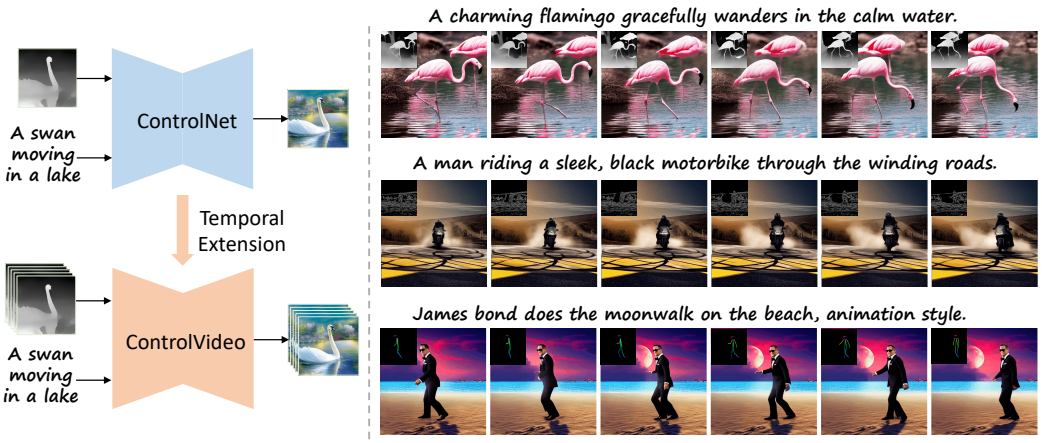

Figure 1: **Training-free controllable text-to-video generation. Left:** ControlVideo adapts ControlNet to the video counterpart by inflating along the temporal axis, aiming to directly inherit its high-quality and consistent generation without any finetuning. **Right:** ControlVideo could synthesize photo-realistic videos conditioned on various motion sequences, which are temporally consistent in both structure and appearance. **Results best seen at 500% zoom.**

(ii), their sparser cross-frame mechanisms increase the discrepancy between the query and key in self-attention modules, and hence impede inheriting high-quality and consistent generation from pre-trained text-to-image models. For (iii), input motion sequences only provide the coarse-level structure of videos, failing to smoothly transition between consecutive frames.

In this work, we propose a training-free *ControlVideo* for high-quality and consistent controllable text-to-video generation, along with *interleaved-frame smoother* to enhance structural smoothness. *ControlVideo* directly inherits the architecture and weights from ControlNet [43], while adapting it to the video counterpart by extending self-attention with the *fully cross-frame interaction*. Different from prior works [15, 40], our fully cross-frame interaction concatenates all frames to become a "larger image", thus directly inheriting high-quality and consistent generation from ControlNet. *Interleaved-frame smoother* deflickers the whole video via the interleaved interpolation at selected sequential timesteps. As illustrated in Fig. 3, the operation at each timestep smooths the interleaved three-frame clips by interpolating middle frames, and the combination at two consecutive timesteps smooths the entire video. Since the smoothing operation is only performed at a few timesteps, the quality and individuality of interpolated frames can be well retained by the following denoising steps.

To enable efficient long-video synthesis, we further introduce a *hierarchical sampler* to produce separated short clips with long-term coherency. In specific, a long video is first split into multiple short video clips with the selected key frames. Then, the key frames are pre-generated with fully cross-frame attention for long-range coherence. Conditioned on pairs of key frames, we sequentially synthesize their corresponding intermediate short video clips with the global consistency.

We conduct the experiments on extensively collected motion-prompt pairs. The experimental results show that our method outperforms alternative competitors qualitatively and quantitatively. Thanks to the efficient designs, *i.e.*, the xFormers [17] implementation and hierarchical sampler, ControlVideo can produce both short and long videos within several minutes using one NVIDIA 2080Ti.

In summary, our contributions are presented as follows:

- We propose a training-free ControlVideo for controllable text-to-video generation, which consists of the fully cross-frame interaction, interleaved-frame smoother, and hierarchical sampler.
- The fully cross-attention demonstrates higher video quality and appearance consistency, while interleaved-frame smoother further reduces structural flickers throughout a whole video.
- The hierarchical sampler enables efficient long-video generation in commodity GPUs.

2 Background

Latent diffusion model (LDM) [29] is an efficient variant of diffusion models [10] by applying the diffusion process in the latent space rather than image space. LDM contains two main components.

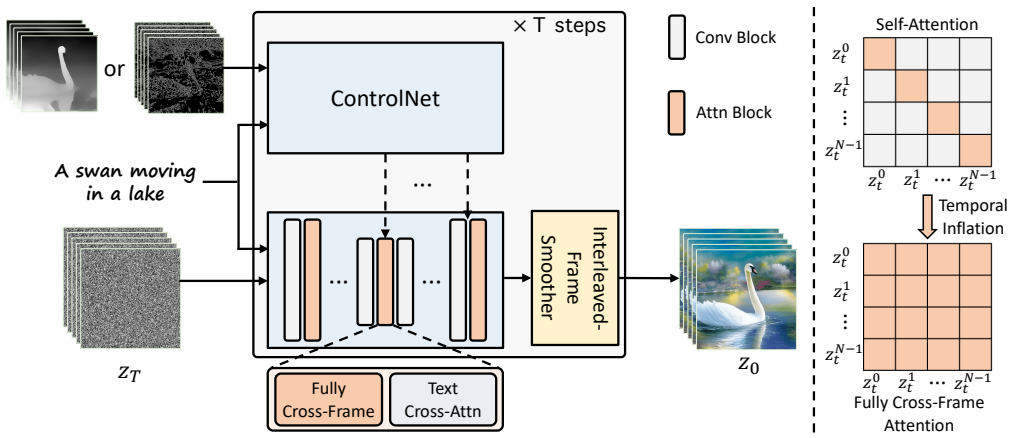

Figure 2: **Overview of ControlVideo.** For consistency in appearance, ControlVideo adapts ControlNet to the video counterpart by adding fully cross-frame interaction into self-attention modules. Considering the flickers in structure, the interleaved-frame smoother is integrated to smooth all inter-frame transitions via the interleaved interpolation (see Fig. 3 for details).

Firstly, it uses an encoder $\mathcal{E}$ to compress an image $\boldsymbol{x}$ into latent code $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$ and a decoder to reconstruct this image $\boldsymbol{x} \approx \mathcal{D}(\boldsymbol{z})$, respectively. Secondly, it learns the distribution of image latent codes $\boldsymbol{z}_0 \sim p_{data}(\boldsymbol{z}_0)$ in a DDPM formulation [10], including a forward and a backward process. The forward diffusion process gradually adds gaussian noise at each timestep t to obtain $\boldsymbol{z}_t$:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t I), \tag{1}$$

where $\{\beta_t\}_{t=1}^T$ are the scale of noises, and T denotes the number of diffusion timesteps. The backward denoising process reverses the above diffusion process to predict less noisy $\boldsymbol{z}_{t-1}$:

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

The μ_θ and Σ_θ are implemented with a denoising model ϵ_θ with learnable parameters θ, which is trained with a simple objective:

$$\mathcal{L}_{simple} := \mathbb{E}_{\mathcal{E}(\boldsymbol{z}),\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t)\|_2^2\Big]. \tag{3}$$

When generating new samples, we start from $\boldsymbol{z}_T \sim \mathcal{N}(0,1)$ and employ DDIM sampling to predict $\boldsymbol{z}_{t-1}$ of previous timestep:

$$\boldsymbol{z}_{t-1} = \sqrt{\alpha_{t-1}}\underbrace{\left(\frac{\boldsymbol{z}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{z}_t, t)}{\sqrt{\alpha_t}}\right)}_{\text{“predicted } \boldsymbol{z}_0\text{”}} + \underbrace{\sqrt{1-\alpha_{t-1}}\cdot\epsilon_\theta(\boldsymbol{z}_t, t)}_{\text{“direction pointing to } \boldsymbol{z}_t\text{”}}, \tag{4}$$

where $\alpha_t = \prod_{i=1}^t(1-\beta_i)$. We use $\boldsymbol{z}_{t\to 0}$ to represent “predicted $\boldsymbol{z}_0$” at timestep t for simplicity. Note that we use Stable Diffusion (SD) $\epsilon_\theta(\boldsymbol{z}_t, t, \tau)$ as our base model, which is an instantiation of text-guided LDMs pre-trained on billions of image-text pairs. τ denotes the text prompt.

ControlNet [43] enables SD to support more controllable input conditions during text-to-image synthesis, *e.g.*, depth maps, poses, edges, *etc*. The ControlNet uses the same U-Net [30] architecture as SD and finetunes its weights to support task-specific conditions, converting $\epsilon_\theta(\boldsymbol{z}_t, t, \tau)$ to $\epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}, \tau)$, where $\boldsymbol{c}$ denotes additional conditions. To distinguish the U-Net architectures of SD and ControlNet, we denote the former as the *main U-Net* while the latter as the *auxiliary U-Net*.

3 ControlVideo

Controllable text-to-video generation aims to produce a video of length N conditioned on motion sequences $\boldsymbol{c} = \{\boldsymbol{c}^i\}_{i=0}^{N-1}$ and a text prompt τ. As illustrated in Fig. 2, we propose a training-free framework termed ControlVideo towards consistent and efficient video generation. Firstly,

ControlVideo is adapted from ControlNet by employing *fully cross-frame interaction*, which ensures the appearance consistency with less quality degradation. Secondly, the *interleaved-frame smoother* deflickers the whole video by interpolating alternate frames at sequential timesteps. Finally, the *hierarchical sampler* separately produces short clips with the holistic coherency to enable long video synthesis.

Fully cross-frame interaction. The main challenge of adapting text-to-image models to the video counterpart is to ensure temporal consistency. Leveraging the controllability of ControlNet, motion sequences could provide coarse-level consistency in structure. Nonetheless, even using the same initial noise, individually producing all frames with ControlNet will lead to drastic inconsistency in appearance (see row 2 in Fig. 6). To keep the video appearance coherent, we concatenate all video frames to become a "large image", so that their content could be shared via inter-frame interaction. Considering that self-attention in SD is driven by appearance similarities [40], we propose to enhance the holistic coherency by adding attention-based fully cross-frame interaction.

In specific, ControlVideo inflates the main U-Net from Stable Diffusion along the temporal axis, while keeping the auxiliary U-Net from ControlNet. Analogous to [11, 15, 40], it directly converts 2D convolution layers to 3D counterpart by replacing 3×3 kernels with $1 \times 3 \times 3$ kernels. In Fig. 2 (right), it extends self-attention by adding interaction across all frames:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}) \cdot \boldsymbol{V}, \text{ where } \boldsymbol{Q} = \boldsymbol{W}^Q \boldsymbol{z}_t,\ \boldsymbol{K} = \boldsymbol{W}^K \boldsymbol{z}_t,\ \boldsymbol{V} = \boldsymbol{W}^V \boldsymbol{z}_t, \quad (5)$$

where $\boldsymbol{z}_t = \{\boldsymbol{z}_t^i\}_{i=0}^{N-1}$ denotes all latent frames at timestep t, while $\boldsymbol{W}^Q$, $\boldsymbol{W}^K$, and $\boldsymbol{W}^V$ project $\boldsymbol{z}_t$ into query, key, and value, respectively.

Previous works [15, 40] usually replace self-attention with sparser cross-frame mechanisms, *e.g.*, all frames attend to the first frame only. Yet, these mechanisms will increase the discrepancy between the query and key in self-attention modules, resulting in the degradation of video quality and consistency. In comparison, our fully cross-frame mechanism combines all frames into a "large image", and has a less generation gap with text-to-image models (see comparisons in Fig. 6). Moreover, with the efficient implementation, the fully cross-frame attention only brings little memory and acceptable computational burden in short-video generation (< 16 frames).

Interleaved-frame smoother. Albeit the videos produced by fully cross-frame interaction are promisingly consistent in appearance, they are still visibly flickering in structure. Input motion sequences only ensure the synthesized videos with coarse-level structural consistency, but not enough to keep the smooth transition between consecutive frames. Therefore, we further propose an interleaved-frame smoother to mitigate the flicker effect in structure. As shown in Fig. 3, our key idea is to smooth each three-frame clip by interpolating the middle frame, following by repeating it in an interleaved manner to smooth the whole video.

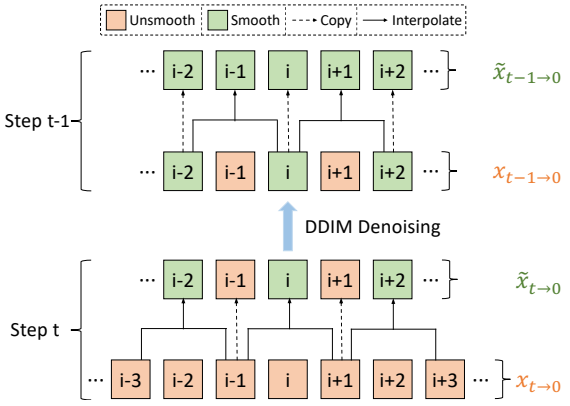

Figure 3: **Illustration of interleaved-frame smoother.** At timestep t, predicted RGB frames $\boldsymbol{x}_{t\to 0}$ are smoothed into $\tilde{\boldsymbol{x}}_{t\to 0}$ via middle-frame interpolation. The combination of two sequential timesteps reduces the structural flickers over the entire video.

Specifically, our interleaved-frame smoother is performed on predicted RGB frames at sequential timesteps. The operation at each timestep interpolates the even or odd frames to smooth their corresponding three-frame clips. In this way, the smoothed three-frame clips from two consecutive timesteps are overlapped together to deflicker the entire video. Before applying our interleaved-frame smoother at timestep t, we first predict the clean video latent $\boldsymbol{z}_{t\to 0}$ according to $\boldsymbol{z}_t$:

$$\boldsymbol{z}_{t\to 0} = \frac{\boldsymbol{z}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}, \tau)}{\sqrt{\alpha_t}}. \quad (6)$$

(a) Depth Maps (b) Canny Edges

Figure 4: **Qualitative comparisons conditioned on depth maps and canny edges.** Our ControlVideo produces videos with better (a) appearance consistency and (b) video quality than others. In contrast, Tune-A-Video [40] fails to inherit structures from source videos, while Text2Video-Zero [15] brings visible artifacts in large motion videos. **Results best seen at 500% zoom.**

After projecting $\boldsymbol{z}_{t\to 0}$ into a RGB video $\boldsymbol{x}_{t\to 0} = \mathcal{D}(\boldsymbol{z}_{t\to 0})$, we convert it to a more smoothed video $\tilde{\boldsymbol{x}}_{t\to 0}$ using our interleaved-frame smoother. Based on smoothed video latent $\tilde{\boldsymbol{z}}_{t\to 0} = \mathcal{E}(\tilde{\boldsymbol{x}}_{t\to 0})$, we compute the less noisy latent $\boldsymbol{z}_{t-1}$ following DDIM denoising in Eq. 4:

$$\boldsymbol{z}_{t-1} = \sqrt{\alpha_{t-1}}\tilde{\boldsymbol{z}}_{t\to 0} + \sqrt{1-\alpha_{t-1}}\cdot\epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}, \tau). \tag{7}$$

Notably, the above process is only performed at the selected intermediate timesteps, which has two advantages: (i) the newly computational burden can be negligible *and* (ii) the individuality and quality of interpolated frames are well retained by the following denoising steps.

Hierarchical sampler. Since video diffusion models need to maintain the temporal consistency with inter-frame interaction, they often require substantial GPU memory and computational resources, especially when producing longer videos. To facilitate efficient and consistent long-video synthesis, we introduce a hierarchical sampler to produce long videos in a clip-by-clip manner. At each timestep, a long video $\boldsymbol{z}_t = \{\boldsymbol{z}_t^i\}_{i=0}^{N-1}$ is separated into multiple short video clips with the selected key frames $\boldsymbol{z}_t^{key} = \{\boldsymbol{z}_t^{kN_c}\}_{k=0}^{\frac{N}{N_c}}$, where each clip is of length $N_c - 1$ and the kth clip is denoted as $\widehat{\boldsymbol{z}}_t^k = \{\boldsymbol{z}_t^j\}_{j=kN_c+1}^{(k+1)N_c-1}$. Then, we pre-generate the key frames with *fully* cross-frame attention for long-range coherence, and their query, key, and value are computed as:

$$\boldsymbol{Q}^{key} = \boldsymbol{W}^Q\boldsymbol{z}_t^{key},\ \boldsymbol{K}^{key} = \boldsymbol{W}^K\boldsymbol{z}_t^{key},\ \boldsymbol{V}^{key} = \boldsymbol{W}^V\boldsymbol{z}_t^{key}. \tag{8}$$

Conditioned on each pair of key frames, we sequentially synthesize their corresponding clips holding the holistic consistency:

$$\widehat{\boldsymbol{Q}}^k = \boldsymbol{W}^Q\widehat{\boldsymbol{z}}_t^k,\quad \widehat{\boldsymbol{K}}^k = \boldsymbol{W}^K[\boldsymbol{z}_t^{kN_c}, \boldsymbol{z}_t^{(k+1)N_c}],\quad \widehat{\boldsymbol{V}}^k = \boldsymbol{W}^V[\boldsymbol{z}_t^{kN_c}, \boldsymbol{z}_t^{(k+1)N_c}]. \tag{9}$$

Table 1: **Quantitative comparisons** of ControlVideo with other methods. We evaluate them on 125 motion-prompt pairs in terms of consistency, and the best results are **bolded**.

Method	Structure Condition	Frame Consistency (%)	Prompt Consistency (%)
Tune-A-Video [40]	DDIM Inversion [35]	94.53	31.57
Text2Video-Zero [15]	Canny Edge	95.17	30.74
ControlVideo	Canny Edge	**96.83**	**30.75**
Text2Video-Zero [15]	Depth Map	95.99	31.69
ControlVideo	Depth Map	**97.22**	**31.81**

4 Experiments

4.1 Experimental Settings

Implementation details. Our ControlVideo is adapted from ControlNet * [43] , and our interleaved-frame smoother employs a lightweight RIFE [13] to interpolate the middle frame of each three-frame clip. The synthesized short videos are of length 15, while the long videos usually contain about 100 frames. Unless otherwise noted, their resolution is both 512×512. During sampling, we adopt DDIM sampling [35] with 50 timesteps, and interleaved-frame smoother is performed on predicted RGB frames at timesteps $\{30, 31\}$ by default. With the efficient implementation of xFormers [17], our ControVideo could produce both short and long videos with one NVIDIA RTX 2080Ti in about 2 and 10 minutes, respectively.

Datasets. To evaluate our ControlVideo, we collect 25 object-centric videos from DAVIS dataset [24] and manually annotate their source descriptions. Then, for each source description, ChatGPT [23] is utilized to generate five editing prompts automatically, resulting in 125 video-prompt pairs in total. Finally, we employ Canny and MiDaS DPT-Hybrid model [28] to estimate the edges and depth maps of source videos, and form 125 motion-prompt pairs as our evaluation dataset. More details are provided in the supplementary materials.

Metrics. Following [5, 40], we adopt CLIP [25] to evaluate the video quality from two perspectives. (i) Frame Consistency: the average cosine similarity between all pairs of consecutive frames, *and* (ii) Prompt Consistency: the average cosine similarity between input prompt and all video frames.

Baselines. We compare our ControlVideo with three publicly available methods: (i) Tune-A-Video [40] extends Stable Diffusion to the video counterpart by finetuning it on a source video. During inference, it uses the DDIM inversion codes of source videos to provide structure guidance. (ii) Text2Video-Zero [15] is based on ControlNet, and employs the first-only cross-frame attention on Stable Diffusion without finetuning. (iii) Follow-Your-Pose [18] is initialized with Stable Diffusion, and is finetuned on LAION-Pose [18] to support human pose conditions. After that, it is trained on millions of videos [41] to enable temporally-consistent video generation.

4.2 Qualitative and quantitative comparisons

Qualitative results. Fig. 4 first illustrates the visual comparisons of synthesized videos conditioned on both (a) depth maps and (b) canny edges. As shown in Fig. 4 (a), our ControlVideo demonstrates better consistency in both appearance and structure than alternative competitors. Tune-A-Video fails to keep the temporal consistency of both appearance and fine-grained structure, *e.g.*, `the color of coat and the structure of road`. With the motion information from depth maps, Text2Video-Zero achieves promising consistency in structure, but still struggles with incoherent appearance in videos *e.g.*, `the color of coat`. Besides, our ControlVideo also performs more robustly when dealing with large motion inputs. As illustrated in Fig. 4 (b), Tune-A-Video ignores the structure information from source videos. Text2Video-Zero adopts the first-only cross-frame mechanism to trade off frame quality and appearance consistency, and generates later frames with visible artifacts. In contrast, with the proposed fully cross-frame mechanism and interleaved-frame smoother, our ControlVideo can handle large motion to generate high-quality and consistent videos.

*https://huggingface.co/lllyasviel/ControlNet

Table 2: **User preference study.** The numbers denote the percentage of raters who favor the videos synthesized by our ControlVideo over other methods.

Method Comparison	Video Quality	Temporal Consistency	Text Alignment
Ours vs. Tune-A-Video [40]	73.6%	83.2%	68.0%
Ours vs. Text2Video-Zero [15]	76.0%	81.6%	65.6%

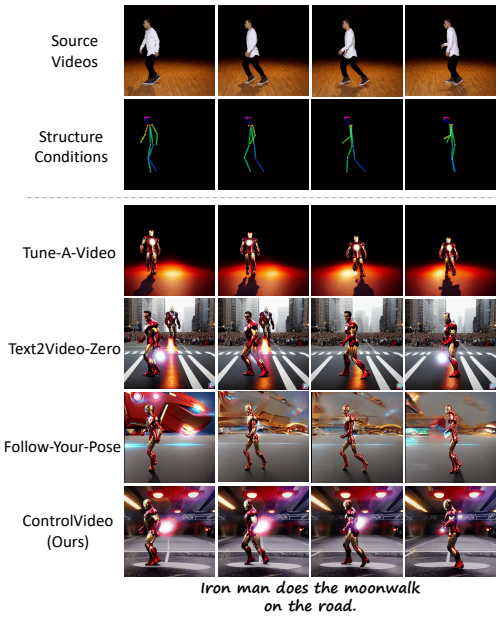

Figure 5: **Qualitative comparisons on poses**. Tune-A-Video [40] only preserves original human positions, while Text2Video-Zero [15] and Follow-Your-Pose [18] produce frames with appearance incoherence, *e.g.*, changing faces of iron man. Our ControlVideo achieves better consistency in both structure and appearance.

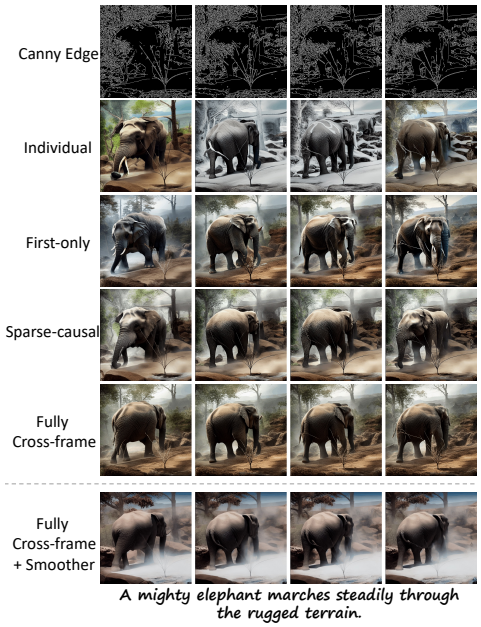

Figure 6: **Qualitative ablation studies** on cross-frame mechanisms and interleaved-frame smoother. Given canny edges in the first row, our fully cross-frame interaction produces video frames with higher quality and consistency than other mechanisms, and adding our smoother further enhances the video smoothness.

Fig. 5 further shows the comparison conditioned on human poses. From Fig. 5, Tune-A-Video only maintains the coarse structures of the source video, *i.e.*, `human position`. Text2Video-Zero and Follow-Your-Pose produce video frames with inconsistent appearance, *e.g.*, `changing faces of iron man` (in row 4) or `disappearing objects in the background` (in row 5). In comparison, our ControlVideo performs more consistent video generation, demonstrating its superiority. More qualitative comparisons are provided in the supplementary materials.

Quantitative results. We have also compared our ControlVideo with existing methods quantitatively on 125 video-prompt pairs. From Table 1, our ControlVideo conditioned on depth outperforms the state-of-the-art methods in terms of frame consistency and prompt consistency, which is consistent with the qualitative results. In contrast, despite finetuning on a source video, Tune-A-Video still struggles to produce temporally coherent videos. Although conditioned on the same structure information, Text2Video-Zero obtains worse frame consistency than our ControlVideo. For each method, the depth-conditioned models generate videos with higher temporal consistency and text fidelity than the canny-condition counterpart, since depth maps provide smoother motion information.

4.3 User study

We then perform the user study to compare our ControlVideo conditioned on depth maps with other competing methods. In specific, we provide each rater a structure sequence, a text prompt, and synthesized videos from two different methods (in random order). Then we ask them to select the better synthesized videos for each of three measurements: (i) video quality, (ii) temporal consistency

Table 3: **Quantitative ablation studies** on cross-frame mechanisms and interleaved-frame smoother. The results indicate that our fully cross-frame mechanism achieves better frame consistency than other mechanisms, and the interleaved-frame smoother significantly improves the frame consistency.

Cross-Frame Mechanism	Frame Consistency (%)	Prompt Consistency (%)	Time Cost (min)
Individual	89.94	30.79	1.2
First-only	94.92	30.54	1.2
Sparse-Causal	95.06	30.59	1.5
Fully	95.36	30.76	3.0
Fully + Smoother	**96.83**	**30.79**	3.5

Figure 7: **A long video produced with our hierarchical sampling.** Motion sequences are shown on the top left. Using the efficient sampler, our ControlVideo generates a high-quality long video with the holistic consistency. **Results best seen at 500% zoom.**

throughout all frames, and (iii) text alignment between prompts and synthesized videos. The evaluation set consists of 125 representative structure-prompt pairs. Each pair is evaluated by 5 raters, and we take a majority vote for the final result. From Table 2, the raters strongly favor our synthesized videos from all three perspectives, especially in temporal consistency. On the other hand, Tune-A-Video fails to generate consistent and high-quality videos with only DDIM inversion for structural guidance, and Text2Video-Zero also produces videos with lower quality and coherency.

4.4 Ablation study

Effect of fully cross-frame interaction. To demonstrate the effectiveness of the fully cross-frame interaction, we conduct a comparison with the following variants: i) individual: no interaction between all frames, ii) first-only: all frames attend to the first one, iii) sparse-causal: each frame attends to the first and former frames, iv) fully: our fully cross-frame, refer to Sec. 3. Note that, all the above models are extended from ControlNet without any finetuning. The qualitative and quantitative results are shown in Fig. 6 and Table 3, respectively. From Fig. 6, the individual cross-frame mechanism suffers from severe temporal inconsistency, *e.g.*, `colorful and black-and-white frames`. The first-only and sparse-causal mechanisms reduce some appearance inconsistency by adding cross-frame interaction. However, they still produce videos with structural inconsistency and visible artifacts, *e.g.*, `the orientation of the elephant and duplicate nose` (row 3 in Fig. 6). In contrast, due to less generation gap with ControlNet, our fully cross-frame interaction performs better appearance coherency and video quality. Though the introduced interaction brings an extra $1 \sim 2\times$ time cost, it is acceptable for a high-quality video generation.

Effect of interleaved-frame smoother. We further analyze the effect of the proposed interleaved-frame smoother. From Fig. 6 and Table 3, our interleaved-frame smoother greatly mitigates structural flickers and improves the video smoothness.

4.5 Extension to long-video generation

Producing a long video usually requires an advanced GPU with high memory. With the proposed hierarchical sampler, our ControlVideo achieves long video generation (more than 100 frames) in a memory-efficient manner. As shown in Fig. 7, our ControlVideo can produce a long video with consistently high quality. Notably, benefiting from our efficient sampling, it only takes approximately ten minutes to generate 100 frames with resolution 512×512 in one NVIDIA RTX 2080Ti. More visualizations of long videos can be found in the supplementary materials.

5 Related work

Text-to-image synthesis. Through pre-training on billions of image-text pairs, large-scale generative models [1, 2, 3, 4, 14, 22, 26, 27, 29, 32, 33, 42] have made remarkable progress in creative and photo-realistic image generation. Various frameworks have been explored to enhance image quality, including GANs [7, 14, 33], autoregressive models [2, 3, 4, 22, 42], and diffusion models [1, 10, 26, 29, 32]. Among these generative models, diffusion-based models are well open-sourced and popularly applied to several downstream tasks, such as image editing [8, 19] and customized generation [6, 16, 31, 37]. Besides text prompts, several works [20, 43] also introduce additional structure conditions to pre-trained text-to-image diffusion models for controllable text-to-image generation. Our ControlVideo is implemented based on the controllable text-to-image models to inherit their ability of high-quality and consistent generation.

Text-to-video synthesis. Large text-to-video generative models usually extend text-to-image models by adding temporal consistency. Earlier works [12, 36, 38, 39] adopt an autoregressive framework to synthesize videos according to given descriptions. Capitalizing on the success of diffusion models in image generation, recent works [9, 11, 34] propose to leverage their potential to produce high-quality videos. Nevertheless, training such large-scale video generative models requires extensive video-text pairs and computational resources.

To reduce the training burden, Gen-1 [5] and Follow-Your-Pose [18] provide coarse temporal information (*e.g.*, motion sequences) for video generation, yet are still costly for most researchers and users. By replacing self-attention with the sparser cross-frame mechanisms, Tune-A-Video [40] and Text2Video-Zero [15] keep considerable consistency in appearance with little finetuning. Our ControlVideo also adapts controllable text-to-image diffusion models without any training, but generates videos with better coherency in both structure and appearance.

6 Discussion

In this paper, we present a training-free framework, namely ControlVideo, towards consistent and efficient controllable text-to-video generation. Particularly, ControlVideo is inflated from ControlNet by adding fully cross-frame interaction to ensure appearance coherence without sacrificing video quality. Besides, interleaved-frame smoother interpolates alternate frames at sequential timesteps to effectively reduce structural flickers. With the further introduced hierarchical sampler and memory-efficient designs, our ControlVideo can generate both short and long videos in several minutes with commodity GPUs. Quantitative and qualitative experiments on extensive motion-prompt pairs demonstrate that ControlVideo performs better than previous state-of-the-arts in terms of video quality and temporal consistency.

Limitations. While our ControlVideo enables consistent and high-quality video generation, it still struggles with producing videos beyond input motion sequences. For example, given sequential poses of `Michael Jackson's moonwalk`, it is difficult to generate a vivid video according to text prompts like `Iron man runs on the street`. In the future, we will explore how motion sequences can be adapted to new ones based on input text prompts, so that users can create more diverse videos with our ControlVideo.

Broader impact. Large-scale diffusion models have made tremendous progress in text-to-video synthesis, yet these models are costly and unavailable to the public. Our ControlVideo focuses on training-free controllable text-to-video generation, and takes an essential step in efficient video creation. Concretely, ControlVideo could synthesize high-quality videos with commodity hardware, hence, being accessible to most researchers and users. For example, artists may leverage our approach to create fascinating videos with less time. Moreover, ControlVideo provides insights into the tasks involved in videoss, *e.g.*, video rendering, video editing, and video-to-video translation. On the flip side, albeit we do not intend to use our model for harmful purposes, it might be misused and bring some potential negative impacts, such as producing deceptive, harmful, or explicit videos. Despite the above concerns, we believe that they could be well minimized with some steps. For example, an NSFW filter can be employed to filter out unhealthy and violent content. Also, we hope that the government could establish and improve relevant regulations to restrict the abuse of video creation.

References

[1] 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. 1, 9

[2] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 9

[3] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *NeurIPS*, 2021. 9

[4] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 9

[5] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 1, 6, 9

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 9

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 9

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 9

[9] 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. 1, 9

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2, 3, 9

[11] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 1, 4, 9

[12] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 1, 9

[13] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *ECCV*, 2022. 6

[14] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023. 9

[15] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 1, 2, 4, 5, 6, 7, 9

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 9

[17] Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, and Daniel Haziza. xformers: A modular and hackable transformer modelling library. https://github.com/facebookresearch/xformers, 2022. 2, 6

[18] 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. 6, 7, 9

[19] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 9

[20] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang 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. 9

[21] Minheng Ni, Zitong Huang, Kailai Feng, and Wangmeng Zuo. Imaginarynet: Learning object detectors without real images and annotations. *arXiv preprint arXiv:2210.06886*, 2022. 1

[22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 9

[23] TB OpenAI. Chatgpt: Optimizing language models for dialogue. *OpenAI*, 2022. 6

[24] 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. 6

[25] 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 *ICML*, 2021. 6

[26] 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. 1, 9

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 9

[28] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 6

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 9

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 9

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 9

[33] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023. 9

[34] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1, 9

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 6

[36] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 9

[37] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 1, 9

[38] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 9

[39] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, 2022. 9

[40] 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. 1, 2, 4, 5, 6, 7, 9

[41] 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 *CVPR*, 2022. 6

[42] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 9

[43] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1, 2, 3, 6, 9