

StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text

Roberto Henschel^{1*}, Levon Khachatryan^{1*}, Daniil Hayrapetyan^{1*}, Hayk Poghosyan¹, Vahram Tadevosyan¹, Zhangyang Wang^{1,2}, Shant Navasardyan¹, Humphrey Shi^{1,3}

¹Picsart AI Resarch (PAIR) ²UT Austin ³SHI Labs @ Georgia Tech, Oregon & UIUC

<https://github.com/Picsart-AI-Research/StreamingT2V>

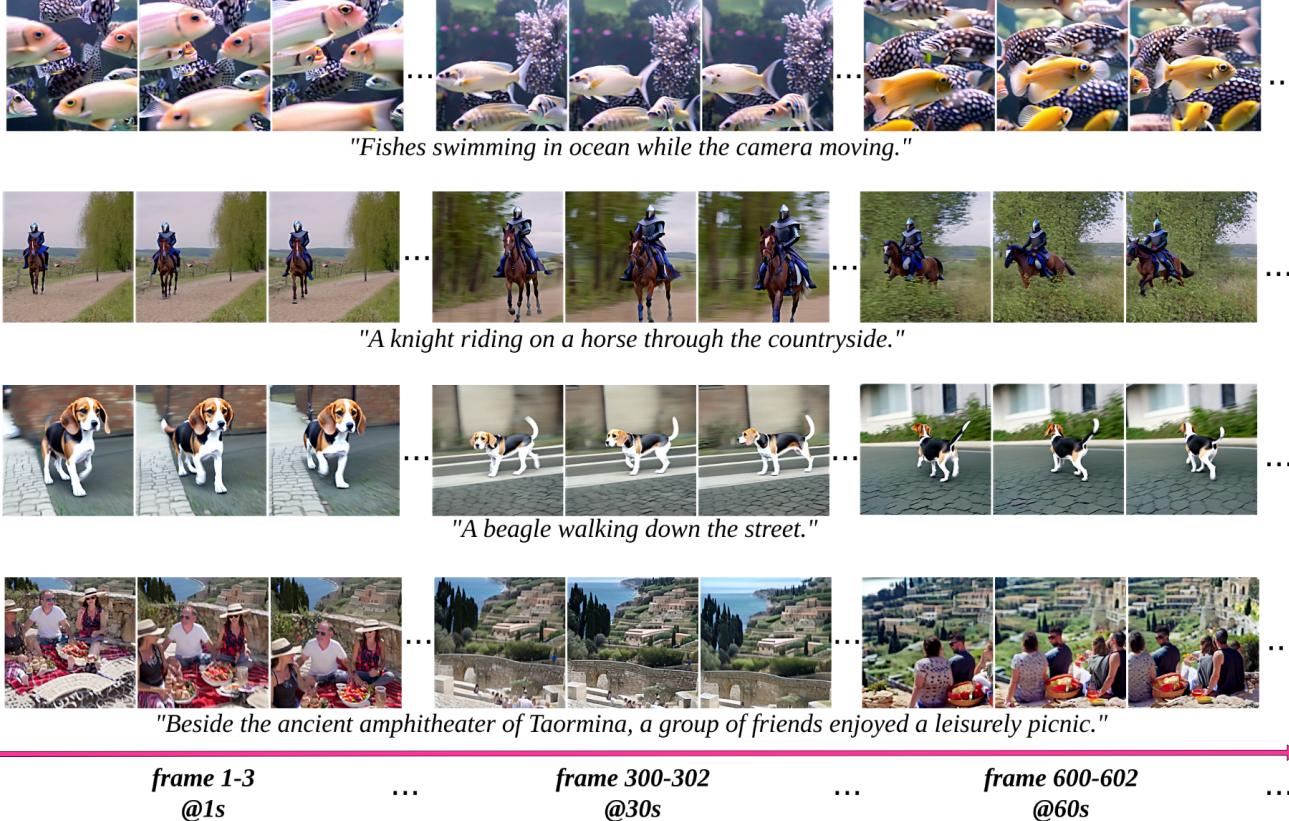


Figure 1. **StreamingT2V** is an advanced autoregressive technique that enables the creation of long videos featuring rich motion dynamics without any stagnation. It ensures temporal consistency throughout the video, aligns closely with the descriptive text, and maintains high frame-level image quality. Our demonstrations include successful examples of videos up to **1200 frames, spanning 2 minutes**, and can be extended for even longer durations. Importantly, the effectiveness of StreamingT2V is not limited by the specific Text2Video model used, indicating that improvements in base models could yield even higher-quality videos.

Abstract

Text-to-video diffusion models enable the generation of high-quality videos that follow text instructions, making it easy to create diverse and individual content. However, ex-

isting approaches mostly focus on high-quality short video generation (typically 16 or 24 frames), ending up with hard-cuts when naively extended to the case of long video synthesis. To overcome these limitations, we introduce StreamingT2V, an autoregressive approach for long video generation of 80, 240, 600, 1200 or more frames with smooth

*Equal contribution.

transitions. The key components are: (i) a short-term memory block called conditional attention module (CAM), which conditions the current generation on the features extracted from the previous chunk via an attentional mechanism, leading to consistent chunk transitions, (ii) a long-term memory block called appearance preservation module, which extracts high-level scene and object features from the first video chunk to prevent the model from forgetting the initial scene, and (iii) a randomized blending approach that enables to apply a video enhancer autoregressively for infinitely long videos without inconsistencies between chunks. Experiments show that StreamingT2V generates high motion amount. In contrast, all competing image-to-video methods are prone to video stagnation when applied naively in an autoregressive manner. Thus, we propose with StreamingT2V a high-quality seamless text-to-long video generator that outperforms competitors with consistency and motion.

1. Introduction

In recent years, with the raise of Diffusion Models [15, 26, 28, 34], the task of text-guided image synthesis and manipulation gained enormous attention from the community. The huge success in image generation led to the further extension of diffusion models to generate videos conditioned by textual prompts [4, 5, 7, 11–13, 17, 18, 20, 32, 37, 39, 45].

Despite the impressive generation quality and text alignment, the majority of existing approaches such as [4, 5, 17, 39, 45] are mostly focused on generating short frame sequences (typically of 16 or 24 frame-length). However, short videos are limited in real-world use-cases such as ad making, storytelling, etc.

The naïve approach of simply training existing methods on long videos (e.g. ≥ 64 frames) is normally unfeasible. Even for generating short sequences, a very expensive training (e.g. using more than 260K steps and 4500 batch size [39]) is typically required. Without training on longer videos, video quality commonly degrades when short video generators are made to output long videos (see appendix). Existing approaches, such as [5, 17, 23], thus extend the baselines to autoregressively generate short video chunks conditioned on the last frame(s) of the previous chunk.

However, the straightforward long-video generation approach of simply concatenating the noisy latents of a video chunk with the last frame(s) of the previous chunk leads to poor conditioning with inconsistent scene transitions (see Sec. 5.3). Some works [4, 8, 40, 43, 48] incorporate also CLIP [25] image embeddings of the last frame of the previous chunk reaching to a slightly better consistency, but are still prone to inconsistent global motion across chunks (see Fig. 5) as the CLIP image encoder loses information important for perfectly reconstructing the conditional frames. The

concurrent work SparseCtrl [12] utilizes a more sophisticated conditioning mechanism by sparse encoder. Its architecture requires to concatenate additional zero-filled frames to the conditioning frames before being plugged into sparse encoder. However, this inconsistency in the input leads to inconsistencies in the output (see Sec. 5.4). Moreover, we observed that all image-to-video methods that we evaluated in our experiments (see Sec. 5.4) lead eventually to video stagnation, when applied autoregressively by conditioning on the last frame of the previous chunk.

To overcome the weaknesses and limitations of current works, we propose **StreamingT2V**, an autoregressive text-to-video method equipped with long/short-term memory blocks that generates long videos without temporal inconsistencies.

To this end, we propose the **Conditional Attention Module (CAM)** which, due to its attentional nature, effectively borrows the content information from the previous frames to generate new ones, while not restricting their motion by the previous structures/shapes. Thanks to CAM, our results are smooth and with artifact-free video chunk transitions.

Existing approaches are not only prone to temporal inconsistencies and video stagnation, but they suffer from object appearance/characteristic changes and video quality degradation over time (see e.g., SVD [4] in Fig. 7). The reason is that, due to conditioning only on the last frame(s) of the previous chunk, they overlook the long-term dependencies of the autoregressive process. To address this issue we design an **Appearance Preservation Module (APM)** that extracts object or global scene appearance information from an initial image (anchor frame), and conditions the video generation process of all chunks with that information, which helps to keep object and scene features across the autoregressive process.

To further improve the quality and resolution of our long video generation, we adapt a video enhancement model for autoregressive generation. For this purpose, we choose a high-resolution text-to-video model and utilize the SDEdit [22] approach for enhancing consecutive 24-frame chunks (overlapping with 8 frames) of our video. To make the chunk enhancement transitions smooth, we design a **randomized blending** approach for seamless blending of overlapping enhanced chunks.

Experiments show that StreamingT2V successfully generates long and temporal consistent videos from text without video stagnation. To summarize, our contributions are three-fold:

- We introduce **StreamingT2V**, an autoregressive approach for seamless synthesis of extended video content using short and long-term dependencies.
- Our **Conditional Attention Module (CAM)** and **Appearance Preservation Module (APM)** ensure the natural continuity of the global scene and object characteristics

of generated videos.

- We seamlessly enhance generated long videos by introducing our ***randomized blending approach*** of consecutive overlapping chunks.

2. Related Work

Text-Guided Video Diffusion Models. Generating videos from textual instructions using Diffusion Models [15, 33] is a recently established yet very active field of research introduced by Video Diffusion Models (VDM) [17]. The approach requires massive training resources and can generate only low-resolution videos (up to 128x128), which are severely limited by the generation of up to 16 frames (without autoregression). Also, the training of text-to-video models usually is done on large datasets such as WebVid-10M [3], or InternVid [41]. Several methods employ video enhancement in the form of spatial/temporal upsampling [5, 16, 17, 32], using cascades with up to 7 enhancer modules [16]. Such an approach produces high-resolution and long videos. Yet, the generated content is still limited by the key frames.

Towards generating longer videos (i.e. more keyframes), Text-To-Video-Zero (T2V0) [18] and ART-V [42] employ a text-to-image diffusion model. Therefore, they can generate only simple motions. T2V0 conditions on its first frame via cross-frame attention and ART-V on an anchor frame. Due to the lack of global reasoning, it leads to unnatural or repetitive motions. MTVG [23] turns a text-to-video model into an autoregressive method by a trainin-free approach. It employs strong consistency priors between and among video chunks, which leads to very low motion amount, and mostly near-static background. FreeNoise [24] samples a small set of noise vectors, re-uses them for the generation of all frames, while temporal attention is performed on local windows. As the employed temporal attentions are invariant to such frame shuffling, it leads to high similarity between frames, almost always static global motion and near-constant videos. Gen-L [38] generates overlapping short videos and aggregates them via temporal co-denoising, which can lead to quality degradations with video stagnation.

Image-Guided Video Diffusion Models as Long Video Generators. Several works condition the video generation by a driving image or video [4, 6–8, 10, 12, 21, 27, 40, 43, 44, 48]. They can thus be turned into an autoregressive method by conditioning on the frame(s) of the previous chunk.

VideoDrafter [21] uses a text-to-image model to obtain an anchor frame. A video diffusion model is conditioned on the driving anchor to generate independently multiple videos that share the same high-level context. However, no consistency among the video chunks is enforced, leading to drastic scene cuts. Several works [7, 8, 44] concatenate

the (encoded) conditionings with an additional mask (which indicates which frame is provided) to the input of the video diffusion model.

In addition to concatenating the conditioning to the input of the diffusion model, several works [4, 40, 48] replace the text embeddings in the cross-attentions of the diffusion model by CLIP [25] image embeddings of the conditional frames. However, according to our experiments, their applicability for long video generation is limited. SVD [4] shows severe quality degradation over time (see Fig. 7), and both, I2VGen-XL[48] and SVD[4] generate often inconsistencies between chunks, still indicating that the conditioning mechanism is too weak.

Some works [6, 43] such as DynamiCrafter-XL [43] thus add to each text cross-attention an image cross-attention, which leads to better quality, but still to frequent inconsistencies between chunks.

The concurrent work SparseCtrl [12] adds a ControlNet [46]-like branch to the model, which takes the conditional frames and a frame-indicating mask as input. It requires by design to append additional frames consisting of black pixels to the conditional frames. This inconsistency is difficult to compensate for the model, leading to frequent and severe scene cuts between frames.

Overall, only a small number of keyframes can currently be generated at once with high quality. While in-between frames can be interpolated, it does not lead to new content. Also, while image-to-video methods can be used autoregressively, their used conditional mechanisms lead either to inconsistencies, or the method suffers from video stagnation. We conclude that existing works are not suitable for high-quality and consistent long video generation without video stagnation.

3. Preliminaries

Diffusion Models. Our text-to-video model, which we term ***StreamingT2V***, is a diffusion model that operates in the latent space of the VQ-GAN [9, 35] autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, where \mathcal{E} and \mathcal{D} are the corresponding encoder and decoder, respectively. Given a video $\mathcal{V} \in \mathbb{R}^{F \times H \times W \times 3}$, composed of F frames with spatial resolution $H \times W$, its *latent code* $x_0 \in \mathbb{R}^{F \times h \times w \times c}$ is obtained through frame-by-frame application of the encoder. More precisely, by identifying each tensor $x \in \mathbb{R}^{F \times \hat{h} \times \hat{w} \times \hat{c}}$ as a sequence $(x^f)_{f=1}^F$ with $x^f \in \mathbb{R}^{\hat{h} \times \hat{w} \times \hat{c}}$, we obtain the latent code via $x_0^f := \mathcal{E}(\mathcal{V}^f)$, for all $f = 1, \dots, F$. The diffusion forward process gradually adds Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to the signal x_0 :

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I), \quad t = 1, \dots, T \quad (1)$$

where $q(x_t | x_{t-1})$ is the conditional density of x_t given x_{t-1} , and $\{\beta_t\}_{t=1}^T$ are hyperparameters. A high value for T is chosen such that the forward process completely destroys

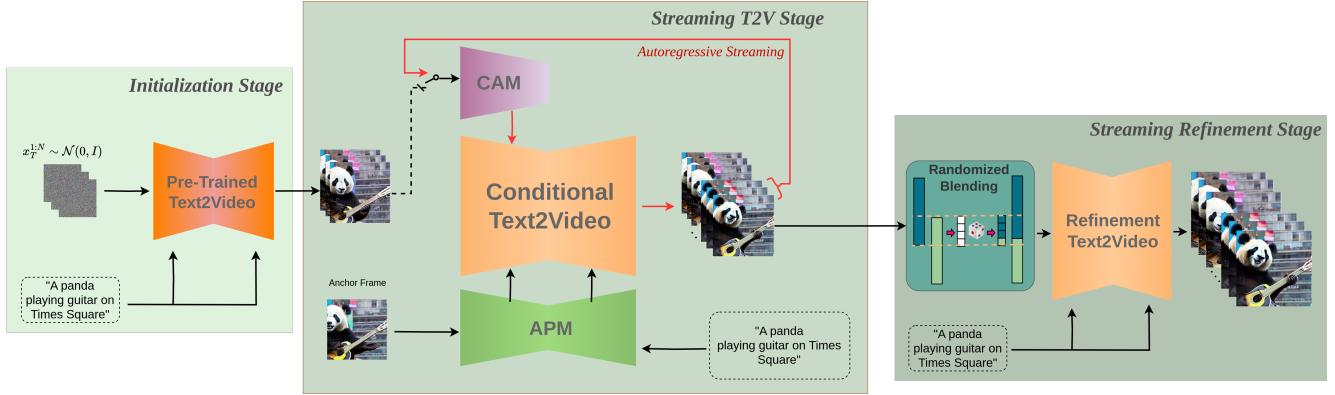


Figure 2. The overall pipeline of StreamingT2V: In the *Initialization Stage* the first 16-frame chunk is synthesized by a text-to-video model (e.g. Modelscope [39]). In the *Streaming T2V Stage* the new content for further frames are autoregressively generated. Finally, in the *Streaming Refinement Stage* the generated long video (600, 1200 frames or more) is autoregressively enhanced by applying a high-resolution text-to-short-video model (e.g. MS-Vid2Vid-XL [48]) equipped with our randomized blending approach.

the initial signal x_0 resulting in $x_T \sim \mathcal{N}(0, I)$. The goal of a diffusion model is then to learn a backward process

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (2)$$

for $t = T, \dots, 1$ (see DDPM [15]), which allows to generate a valid signal x_0 from standard Gaussian noise x_T . Once x_0 is obtained from x_T , we obtain the generated video through frame-wise application of the decoder: $\tilde{\mathbf{V}}^f := \mathcal{D}(x_0^f)$, for all $f = 1, \dots, F$. Yet, instead of learning a predictor for mean and variance in Eq. (2), we learn a model $\epsilon_\theta(x_t, t)$ to predict the Gaussian noise ϵ that was used to form x_t from input signal x_0 (which is a common reparametrization [15]).

To guide the video generation by a textual prompt τ , we use a noise predictor $\epsilon_\theta(x_t, t, \tau)$ that is conditioned on τ . We model $\epsilon_\theta(x_t, t, \tau)$ as a neural network with learnable weights θ and train it on the denoising task:

$$\min_{\theta} \mathbb{E}_{t, (x_0, \tau) \sim p_{data}, \epsilon \sim \mathcal{N}(0, I)} \|\epsilon - \epsilon_\theta(x_t, t, \tau)\|_2^2, \quad (3)$$

using the data distribution p_{data} . To simplify notation, we will denote by $x_t^{r:s} = (x_t^j)_{j=r}^s$ the latent sequence of x_t from frame r to frame s , for all $r, t, s \in \mathbb{N}$.

Text-To-Video Models. Most text-to-video models [5, 11, 16, 32, 39] extend pre-trained text-to-image models [26, 28] by inserting new layers that operate on the temporal axis. Modelscope (MS) [39] follows this approach by extending the UNet-like [29] architecture of Stable Diffusion [28] with temporal convolutional and attentional layers. It was trained in a large-scale setup to generate videos with 3 FPS@256x256 and 16 frames. The quadratic growth in memory and compute of the temporal attention layers (used in recent text-to-video models) together with very high training costs limits current text-to-video models to generate long sequences. In this paper, we demonstrate our

StreamingT2V method by taking MS as a basis and turn it into an autoregressive model suitable for long video generation with high motion dynamics and consistency.

4. Method

In this section, we introduce our method for high-resolution text-to-long video generation. We first generate 256×256 resolution long videos for 5 seconds (16fps), then enhance them to higher resolution (720×720). The overview of the whole pipeline is provided in Fig. 2. The long video generation part consists of (*Initialization Stage*) synthesizing the first 16-frame chunk by a pre-trained text-to-video model (for example one may take Modelscope [39]), and (*Streaming T2V Stage*) autoregressively generating the new content for further frames. For the autoregression (see Fig. 3), we propose our conditional attention module (CAM) that leverages short-term information from the last $F_{cond} = 8$ frames of the previous chunk to enable seamless transitions between chunks. Furthermore, we leverage our appearance preservation module (APM), which extracts long-term information from a fixed anchor frame making the autoregression process robust against losing object appearance or scene details in the generation.

After having a long video (80, 240, 600, 1200 frames or more) generated, we apply the *Streaming Refinement Stage* which enhances the video by autoregressively applying a high-resolution text-to-short-video model (for example one may take MS-Vid2Vid-XL [48]) equipped by our randomized blending approach for seamless chunk processing. The latter step is done without additional training by so making our approach affordable with lower computational costs.

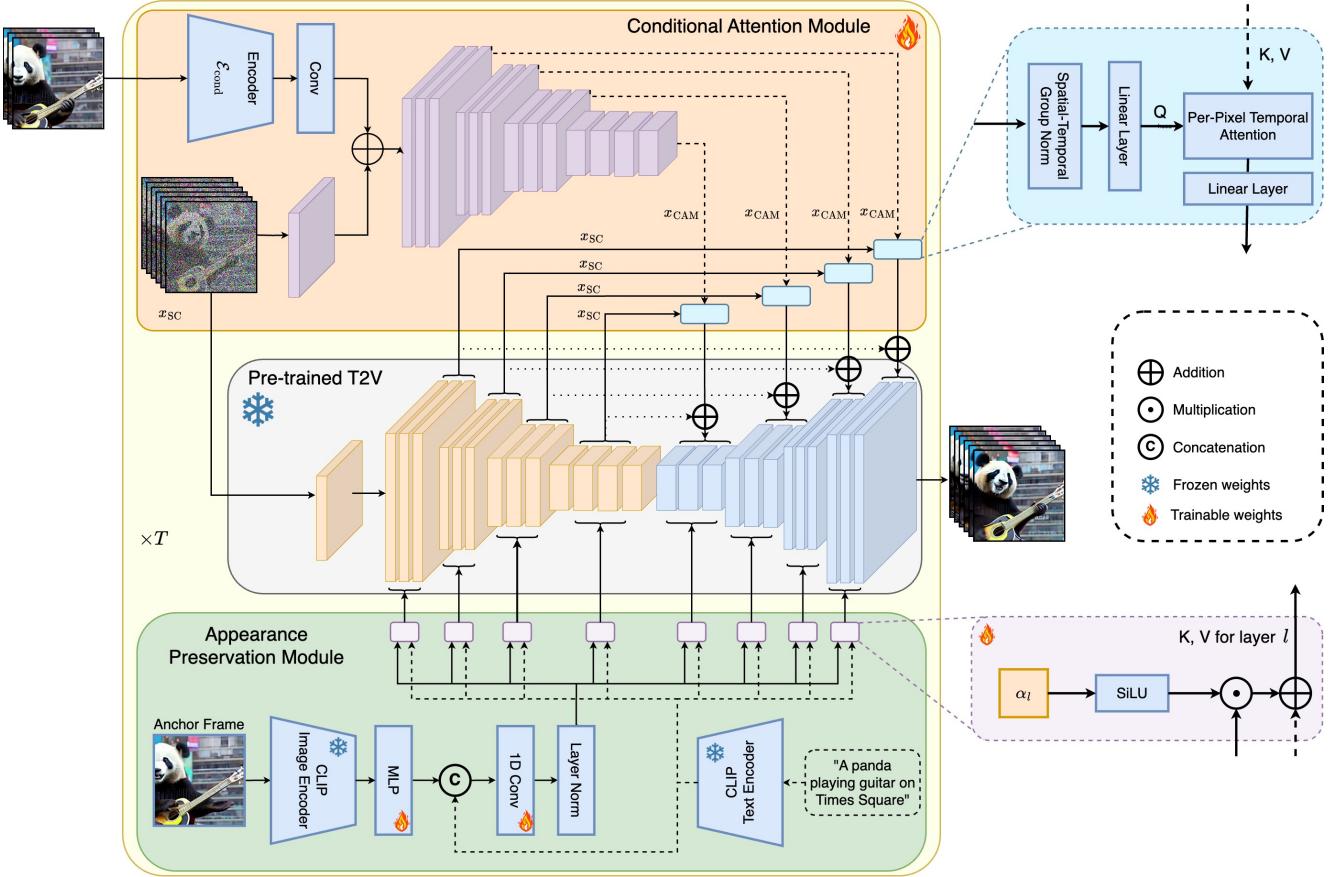


Figure 3. Method overview: StreamingT2V extends a video diffusion model (VDM) by the conditional attention module (CAM) as short-term memory, and the appearance preservation module (APM) as long-term memory. CAM conditions the VDM on the previous chunk using a frame encoder $\mathcal{E}_{\text{cond}}$. The attentional mechanism of CAM leads to smooth transitions between chunks and videos with high motion amount at the same time. APM extracts from an anchor frame high-level image features and injects it to the text cross-attentions of the VDM. APM helps to preserve object/scene features across the autoregressive video generation.

4.1. Conditional Attention Module

To train a conditional network for our Streaming T2V stage, we leverage the pre-trained power of a text-to-video model (e.g. Modelscope [39]), as a prior for long video generation in an autoregressive manner. In the further writing we will refer this pre-trained text-to-(short)video model as **Video-LDM**. To autoregressively condition Video-LDM by some short-term information from the previous chunk (see Fig. 2, mid), we propose the **Conditional Attention Module (CAM)**, which consists of a feature extractor, and a feature injector into Video-LDM UNet, inspired by ControlNet [46]. The feature extractor utilizes a frame-wise image encoder $\mathcal{E}_{\text{cond}}$, followed by the same encoder layers that the Video-LDM UNet uses up to its middle layer (and initialized with the UNet’s weights). For the feature injection, we let each long-range skip connection in the UNet *attend* to corresponding features generated by CAM via cross-attention.

Let x denote the output of $\mathcal{E}_{\text{cond}}$ after zero-convolution. We use addition, to fuse x with the output of the first temporal transformer block of CAM. For the injection of CAM’s features into the Video-LDM Unet, we consider the UNet’s skip-connection features $x_{\text{SC}} \in \mathbb{R}^{b \times F \times h \times w \times c}$ (see Fig.3) with batch size b . We apply spatio-temporal group norm, and a linear projection P_{in} on x_{SC} . Let $x'_{\text{SC}} \in \mathbb{R}^{(b \cdot w \cdot h) \times F \times c}$ be the resulting tensor after reshaping. We condition x'_{SC} on the corresponding CAM feature $x_{\text{CAM}} \in \mathbb{R}^{(b \cdot w \cdot h) \times F_{\text{cond}} \times c}$ (see Fig.3), where F_{cond} is the number of conditioning frames, via temporal multi-head attention (T-MHA) [36], i.e. independently for each spatial position (and batch). Using learnable linear maps P_Q, P_K, P_V , for queries, keys, and values, we apply T-MHA using keys and values from x_{CAM} and queries from

x'_{SC} , i.e.

$$\begin{aligned} x''_{\text{SC}} &= \text{T-MHA}(Q = P_Q(x'_{\text{SC}}), \\ &\quad K = P_K(x_{\text{CAM}}), \\ &\quad V = P_V(x_{\text{CAM}})). \end{aligned} \quad (4)$$

Finally, we use a linear projection P_{out} . Using a suitable reshaping operation R , the output of CAM is added to the skip connection (as in ControlNet [46]):

$$x'''_{\text{SC}} = x_{\text{SC}} + R(P_{\text{out}}(x''_{\text{SC}})), \quad (5)$$

so that x''' is used in the decoder layers of the UNet. The projection P_{out} is zero-initialized, so that when training starts, CAM is not affecting the base model’s output, which improves training convergence.

CAM utilizes the last F_{cond} conditional frames of the previous chunk as input. The cross attention enables to condition the F frames of the base model to CAM. In contrast, the sparse encoder [12] uses convolution for the feature injection, and thus needs additional $F - F_{\text{cond}}$ zero-valued frames (and a mask) as input in order to add the output to the F frames of the base model. This poses an inconsistency in the input for SparseCtrl, leading to severe inconsistencies in the generated videos (see Sect. 5.3 and Sect. 5.4).

4.2. Appearance Preservation Module

Autoregressive video generators typically suffer from forgetting initial object and scene features, leading to severe appearance changes. To tackle this issue, we incorporate long-term memory by leveraging the information contained in a fixed anchor frame of the very first chunk using our proposed **Appearance Preservation Module (APM)**. This helps to maintain scene and object features across video chunk generations (see Fig. 6).

To enable APM to balance the guidance by the anchor frame with the guidance by the text instructions, we propose (see Figure 3): (i) We mix the CLIP [25] image token of the anchor frame with the CLIP text tokens from the textual instruction by expanding the clip image token to $k = 8$ tokens using a linear layer, concatenate the text and image encodings at the token dimension, and use a projection block, leading to $x_{\text{mixed}} \in \mathbb{R}^{b \times 77 \times 1024}$; (ii) We introduce for each cross attention layer l a weight $\alpha_l \in \mathbb{R}$ (initialized as 0) to perform cross attention using keys and values coming from a weighted sum of x_{mixed} and the usual CLIP text encoding of the text instructions x_{text} :

$$x_{\text{cross}} = \text{SiLU}(\alpha_l)x_{\text{mixed}} + x_{\text{text}}. \quad (6)$$

The experiments in Section 5.3 show that the lightweight APM module helps to keep scene and identity features across the autoregressive process (see Fig. 6).

4.3. Auto-regressive Video Enhancement

To further improve quality and resolution of our text-to-video results, we utilize a high-resolution (1280x720) text-to-(short)video model (Refiner Video-LDM, see Fig. 3) to autoregressively enhance 24-frame chunks of generated videos. Using a text-to-video model as a refiner/enhancer of 24-frame chunks is done by adding a substantial amount of noise to the input video-chunk and denoising with the text-to-video diffusion model (SDEdit [22] approach). More precisely, we take a high-resolution text-to-video model (for example MS-Vid2Vid-XL [40, 48]), and a low resolution video chunk of 24 frames which is first bilinearly upscaled [2] to the target high resolution. Then we encode the frames using the image encoder \mathcal{E} so that we obtain a latent code x_0 . After that we apply $T' < T$ forward diffusion steps (see Eq. (1)) so that $x_{T'}$ still contains signal information (mostly about the video structure), and denoise it using the high-resolution video diffusion model.

However, the naïve approach of independently enhancing each chunk leads to inconsistent transitions (see Fig. 4 (a)). We tackle this issue by using shared noise between consecutive chunks and leveraging our randomized blending approach. Given our low-resolution long video, we split it into m chunks $\mathcal{V}_1, \dots, \mathcal{V}_m$ of $F = 24$ frame-length such that each two consecutive chunks have an overlap of $\mathcal{O} = 8$ frames. For the backward diffusion at step t , starting from T' , we must sample noise to perform one denoising step (see Eq. 2). We start with the first chunk \mathcal{V}_1 and sample noise $\epsilon_1 \sim \mathcal{N}(0, I)$ with $\epsilon_1 \in \mathbb{R}^{F \times h \times w \times c}$. For each subsequent chunk \mathcal{V}_i , $i > 1$, we sample noise $\hat{\epsilon}_i \sim \mathcal{N}(0, I)$ with $\hat{\epsilon}_i \in \mathbb{R}^{(F-\mathcal{O}) \times h \times w \times c}$ and concatenate it along the frame dimension with the noise $\epsilon_{i-1}^{F-\mathcal{O}:F}$ that was sampled for the \mathcal{O} overlapping frames of the previous chunk, i.e.

$$\epsilon_i := \text{concat}([\epsilon_{i-1}^{F-\mathcal{O}:F}, \hat{\epsilon}_i], \text{dim} = 0), \text{ for all } i = 2, \dots, m, \quad (7)$$

so that we obtain shared noise for overlapping frames. We perform one denoising step using ϵ_i and obtain for chunk \mathcal{V}_i the latent code $x_{t-1}(i)$. Yet, this approach is not sufficient to remove transition misalignment (see Fig. 4 (b)).

To improve consistency significantly, we propose the **randomized blending** approach. Consider the latent codes $x_{t-1}^{F-\mathcal{O}:F}(i-1)$ and $x_{t-1}^{1:\mathcal{O}}(i)$ of two consecutive chunks $\mathcal{V}_{i-1}, \mathcal{V}_i$ at denoising step $t-1$. The latent code $x_{t-1}(i-1)$ of chunk \mathcal{V}_{i-1} possesses a smooth transition from its first frames to the overlapping frames, while the latent code $x_{t-1}(i)$ possesses a smooth transition from the overlapping frames to the subsequent frames. Thus, we combine the two latent codes via concatenation, by randomly sampling a frame index f_{thr} from $\{0, \dots, \mathcal{O}\}$ then taking from $x_{t-1}^{F-\mathcal{O}:F}(i-1)$ the latent code of the first f_{thr} frames and from $x_{t-1}^{1:\mathcal{O}}(i)$ the latent code of the frames starting from $f_{\text{thr}} + 1$. Then, we update the latent code of the entire long

video x_{t-1} on the overlapping frames and perform the next denoising step. Accordingly, for a frame $f \in \{1, \dots, \mathcal{O}\}$ of the overlap and diffusion step t , the latent code of chunk \mathcal{V}_{i-1} is used with probability $1 - \frac{f}{\mathcal{O}+1}$.

By using a probabilistic mixture of the latents in an overlapping region, we successfully diminish inconsistencies between chunks (see Fig. 4(c)).

5. Experiments

5.1. Implementation Details

We generate $F = 16$ frames and condition on $F_{\text{cond}} = 8$ frames. Training is conducted using a dataset collected from publicly available sources. We sample with 3FPS@256x256 16 frames (during CAM training) and 32 frames (during CAM+APM training). First, we freeze the weights of the pre-trained Video-LDM and train the new layers of CAM with batch size 8 and learning rate $5 \cdot 10^{-5}$ for 400K steps. We then continue to train CAM+APM. We randomly sample an anchor frame out of the first 16 frames.

For the conditioning and denoising, we use the first 8 and 16 frames, starting from frame 17, respectively. This aligns training with inference, where there is a large time gap between the conditional frames and the anchor frame. In addition, by randomly sampling an anchor frame, the model can leverage the CLIP information only for the extraction of high-level semantic information, as we do not provide a frame index to the model. We freeze the CLIP encoder and the temporal layers of the main branch, and train the remaining layers for 1K steps.

The image encoder $\mathcal{E}_{\text{cond}}$ used in CAM is composed of stacked 2D convolutions, layer norms and SiLU activations. For the video enhancer, we diffuse an input video using $T' = 600$ steps. Additional implementation details of StreamingT2V are provided in the appendix.

5.2. Metrics

For quantitative evaluation we employ metrics that measure to evaluate temporal consistency, text-alignment, and per-frame quality of our method.

For temporal consistency, we introduce SCuts, which counts for a video the number of detected scene cuts, using the AdaptiveDetector algorithm of the PySceneDetect [1] package with default parameters. In addition, we propose a new metric called **motion aware warp error (MAWE)**, which coherently assesses motion amount and warp error, and yields a low value when the video exhibits both consistency and a substantial amount of motion. To this end, we measure the motion amount using OFS (optical flow score), which computes for a video the mean magnitude of all optical flow vectors between any two consecutive frames. Furthermore, for a video \mathcal{V} , we consider the mean warp error [19] $W(\mathcal{V})$, which measures the average squared L2 pixel

distance from a frame to its warped subsequent frame, excluding occluded regions. Finally, MAWE is defined as:

$$\text{MAWE}(\mathcal{V}) := \frac{W(\mathcal{V})}{c \cdot \text{OFS}(\mathcal{V})}, \quad (8)$$

where c aligns the different scales of the two metrics. To this end, we perform a regression analysis on a subset of our dataset validation videos and obtain $c = 9.5$ (details to derive c are provided in the appendix). MAWE requires high motion amount *and* low warp error for a low metric values. For the metrics involving optical flow, computations are conducted by resizing all videos to 720×720 resolution.

For video textual alignment, we employ the CLIP [25] text image similarity score (CLIP), which is applied to all frames of a video. CLIP computes for a video sequence the cosine similarity from the CLIP text encoding to the CLIP image encodings.

For per-frame quality we incorporate the Aesthetic Score [31] computed on top of CLIP image embeddings of all frames of a video.

All metrics are computed per video first and then averaged over all videos, all videos are generated with 80 frames for quantitative analysis.

5.3. Ablation Study

To assess our proposed components, we perform ablation studies on a randomly sampled set of 75 prompts from our validation set. We compare CAM against established conditioning approaches, analyse the impact of our long-term memory APM, and ablate on our modifications for the video enhancer.

CAM for Conditioning. To analyse the importance of CAM, we compare CAM (w/o APM) with two baselines: (i) Connect the features of CAM with the skip-connection of the UNet via zero convolution, followed by addition. We zero-pad the condition frame and concatenate it with an frame-indicating mask to form the input for the modified CAM, which we denote as Add-Cond. (ii) We append the conditional frames and a frame-indicating mask to input of Video-LDM’s Unet along the channel dimension, but do not use CAM, which we denote as Conc-Cond. We train our method with CAM and the baselines on the same dataset. Architectural details (including training) of these baselines are provided in the appendix.

We obtain an SCuts score of 0.24, 0.284 and 0.03 for Conc-Cond, Add-Cond and Ours (w/o APM), respectively. This shows that the inconsistencies in the input caused by the masking leads to frequent inconsistencies in the generated videos and that concatenation to the Unet’s input is a too weak conditioning. In contrast, our CAM generates consistent videos with a SCuts score that is 88% lower than the baselines.

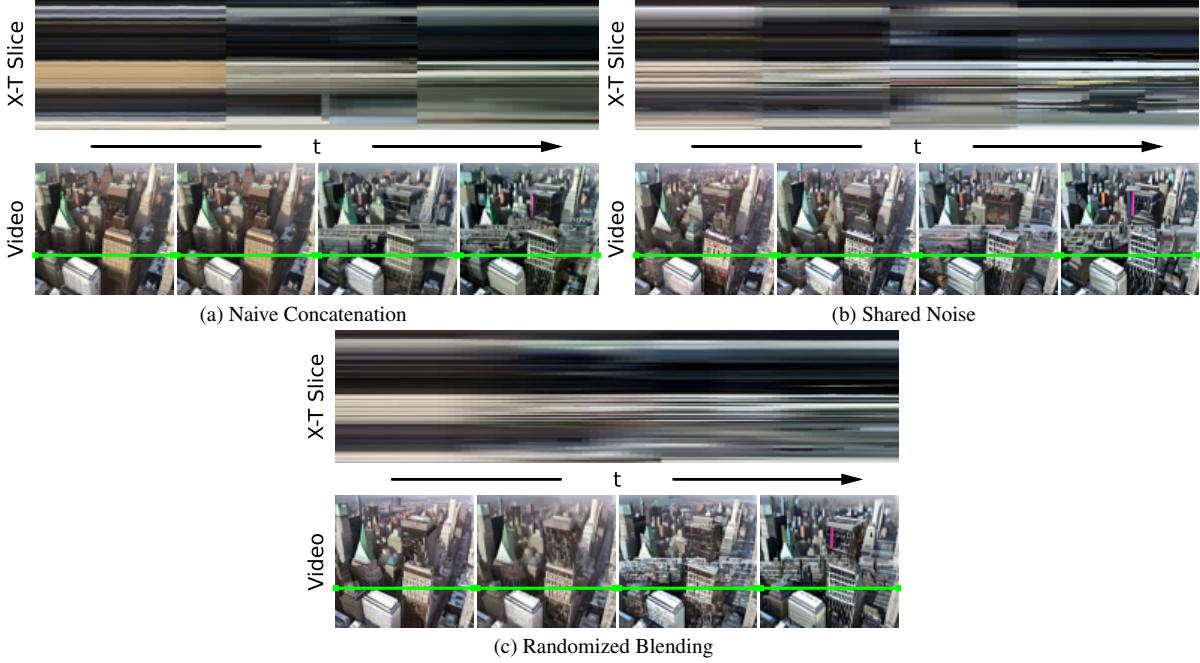


Figure 4. Ablation study on our video enhancer improvements. The X-T slice visualization shows that randomized blending leads to smooth chunk transitions, while both baselines have clearly visible, severe inconsistencies between chunks.

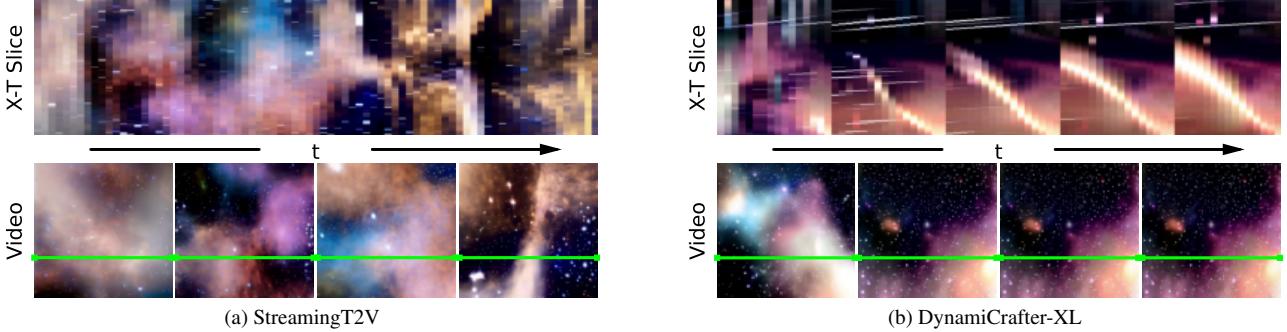


Figure 5. Visual comparision of DynamiCrafter-XL and StreamingT2V. Both text-to-video results are generated using the same prompt. The X-T slice visualization shows that DynamiCrafter-XL suffers from severe chunk inconsistencies and repetitive motions. In contrast, our method shows seamless transitions and evolving content.

Long-Term Memory. We analyse the impact of utilizing long-term memory in the context of long video generation.

Fig. 6 shows that long-term memory greatly helps keeping the object and scene features across autoregressive generations. This is also supported quantitatively. We obtain a person re-identification score (definition in the appendix) of 93.42 and 94.95 for Ours w/o APM, and Ours, respectively. Our APM module thus improves the identity/appearance preservation. Also the scene information is better kept, as we observe an image distance score in terms of LPIPS[47] of 0.192 and 0.151 for Ours w/o APM and Ours, respectively. We thus have an improvement in terms of scene preservation of more than 20% when APM is used.

Randomized Blending for Video Enhancement. We assess our randomized blending approach by comparing against two baselines. (B) enhances each video chunk independently, and (B+S) uses shared noise for consecutive chunks, with an overlap of 8 frames, but not randomized blending. We compute per sequence the standard deviation of the optical flow magnitudes between consecutive frames and average over all frames and sequences, which indicates temporal smoothness. We obtain the scores 8.72, 6.01 and 3.32 for B, B+S, and StreamingT2V, respectively. Thus, noise sharing improves chunk consistency (by 31% vs B), but it is significantly further improved by randomized blending (by 62% vs B). Also the visual results in Fig.



(a) Young caucasian female couple drinking cocktails and smiling on terrace in havana, cuba. girls, teens, teenagers, women

Figure 6. Top row: CAM+APM, Bottom row: CAM. The figure shows that using long-term information via APM helps to keep identities (e.g. the face of the left woman) and scene features, e.g. the dresses or arm clock.

4, where only StreamingT2V is able to produce seamless transitions, confirms the benefit of randomized blending.

5.4. Comparison with Baselines

Benchmark. To assess the effectiveness of StreamingT2V, we create a test set composed of 50 prompts covering different actions, objects and scenes (prompts are listed in the appendix). We compare against recent methods that have code available, namely the image-to-video methods I2VGen-XL [48], SVD [4], DynamiCrafter-XL [43] and SEINE [7] used autoregressively, the video-to-video method SparseControl [12], and the text-to-long-video method FreeNoise [24].

For all methods, we use their released model weights and hyperparameters. To have a fair comparison and insightful analysis on the performance of the methods for the autoregressive generation, independent of the employed initial frame generator, we use the same Video-LDM model to generate the first chunk consisting of 16 frames, given a text prompt and enhance it to 720x720 resolution using the same Refiner Video-LDM. Then, we generate the videos, while we start all autoregressive methods by conditioning on the last frame(s) of that chunk. For methods working on different spatial resolution, we apply zero padding to the first chunk.

Automatic Evaluation. Our quantitative evaluation on the test set shows that StreamingT2V performs best in terms of seamless chunk transitions and motion consistency (see Tab. 8). Our MAWE score is significantly better than for all competing methods (e.g. more than 50% lower than the second best score by SEINE). This is also supported by the lowest SCuts score among all competitors. Only the methods FreeNoise and SEINE achieve the same score. However, they produce near-static videos (see also Fig. 7), leading automatically to low SCuts scores. While SparseControl also follows a ControlNet approach, it leads to 50 times more scene cuts compared to StreamingT2V. This shows the advantage of our attentional CAM block over SparseControl, where the conditional frames need to be pad with zeros, so that inconsistency in the input lead to severe scene

cuts.

Interestingly, all competing methods that incorporate CLIP image encodings are prone to misalignment (measured in low CLIP scores), i.e. SVD and DynamiCrafterXL and I2VGen-XL.

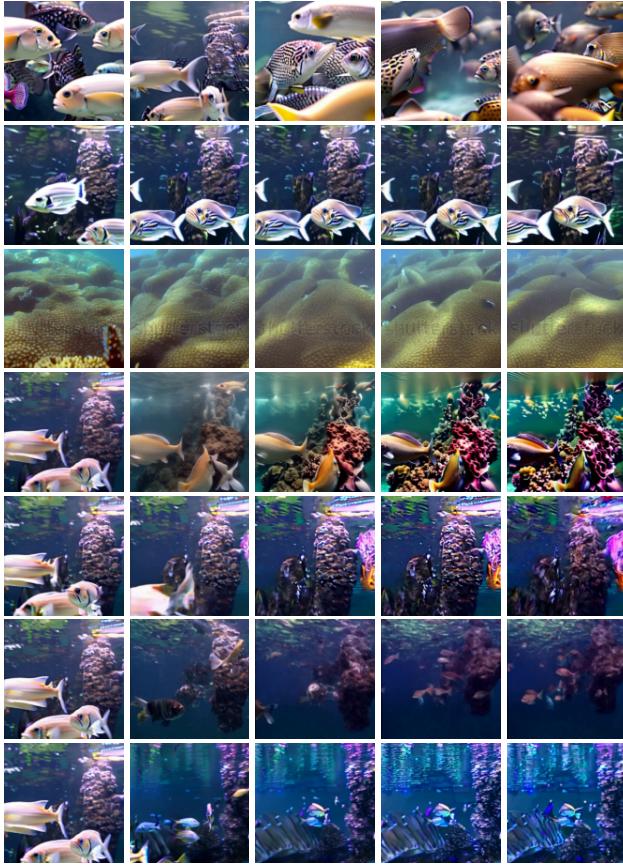
We hypothesize that this is due to a domain shift; the CLIP image encoder is trained on natural images, but applied on autoregressively generated images. With the help of our long-term memory, APM reminds the network about the domain of real images, as we use a fixed anchor frame, so that it does not degrade, leading to the second highest CLIP score.

Moreover, we evaluated the per-frame quality of our generated videos, and, as can be noticed from Tab. 8 StreamingT2V significantly outperforms almost all the competitors and slightly underperform SparseCtrl. This shows that our method is able to produce high-quality long videos with much temporal consistency and motion dynamics than the competitors.

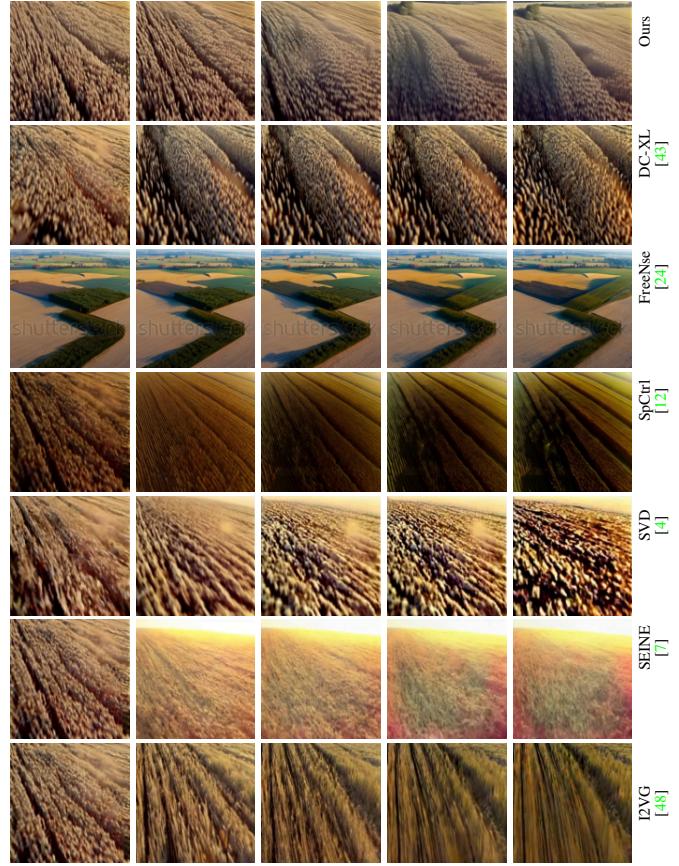
Qualitative Evaluation. Finally, we present corresponding visual results on the test set in Fig. 7. The high similarity of the frames depicted in middle shows that all competing methods suffer from video stagnation, where the background and the camera is freezed and nearly no location motion is generated. In contrast, our method is generating smooth and consistent videos without leading to standstill. SVD, SparseCtrl and SEINE are prone to severe quality degradation, e.g., wrong colors and distorted frames and DynamiCrafter-XL shows strong inconsistencies between chunks (see Fig. 5), showing that their conditioning via CLIP image encoder and concatenation is too weak. Thanks to the more powerful CAM mechanism, StreamingT2V leads to smooth chunk transitions.

6. Conclusion

In this paper, we addressed long video generation from textual prompts. We identified that all existing methods are failing to solve this task satisfactorily, generating either



(a) Fishes swimming in ocean camera moving, cinematic



(b) Close flyover over a large wheat field in the early morning sunlight, cinematic

Figure 7. Visual comparisons of StreamingT2V with state-of-the-art methods on 80 frame-length, autoregressively generated videos. In contrast to other methods, StreamingT2V generates long videos without suffering from motion stagnation.

Table 8. Quantitative comparison to state-of-the-art open-source text-to-long-video generators. Best performing metrics are highlighted in red, second best in blue. We also tested OpenAI’s Sora with their 48 sample videos released on their website. We list the numbers here to give the readers a sense of how Sora is performing on these metrics but please be advised that this test is different from the other open-source models both in terms of the test set and prompts in this table, and hence it is in gray and not ranked.

Method	\downarrow MAWE	\downarrow SCuts	\uparrow CLIP	\uparrow AE
	More motion dynamics / less video stagnation	Better video consistency / less scene change	Better text alignment	Better image quality
SparseCtrl [12]	39.92	2.02	31.13	5.29
I2VGenXL [48]	72.89	0.44	29.85	4.71
DynamiCrafterXL [43]	28.60	1.76	30.50	4.83
SEINE [7]	23.69	0.04	31.04	4.71
SVD [4]	96.17	1.34	28.89	3.90
FreeNoise [24]	49.53	0.04	32.14	4.79
OpenAI Sora (<i>demo-only</i>)	9.26	0.81	33.01	5.42
StreamingT2V (<i>Ours</i>)	10.87	0.04	31.30	5.07

videos with temporal inconsistencies, or severe stagnation up to standstill. To overcome these limitations, we carefully

analysed an autoregressive pipeline build on top of a vanilla video diffusion model and proposed StreamingT2V, which

utilizes novel short- and long-term dependency blocks for seamless continuation of video chunks with a high motion amount, while preserving high-level scene and object features throughout the generation process. We proposed a randomized blending approach that enables to use a video enhancer within the autoregressive process without temporal inconsistencies. Experiments showed that StreamingT2V leads to significant higher motion amount and temporal consistency compared with all competitors. StreamingT2V enables long video generation from text without stagnating content.

References

- [1] Pyscenedetect. <https://www.scenedetect.com/>. Accessed: 2024-03-03. 7
- [2] Isaac Amidror. Scattered data interpolation methods for electronic imaging systems: a survey. *Journal of electronic imaging*, 11(2):157–176, 2002. 6
- [3] Max Bain, Arsha Nagrani, GÜl Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 3
- [4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3, 9, 10
- [5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 3, 4
- [6] Xi Chen, Zhiheng Liu, Mengting Chen, Yutong Feng, Yu Liu, Yujun Shen, and Hengshuang Zhao. Livephoto: Real image animation with text-guided motion control. *arXiv preprint arXiv:2312.02928*, 2023. 3
- [7] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 3, 9, 10
- [8] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Animateanything: Fine-grained open domain image animation with motion guidance, 2023. 2, 3
- [9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3
- [10] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 3, 15
- [11] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 2, 4
- [12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023. 2, 3, 6, 9, 10
- [13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2023. 2
- [14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 15
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 4
- [16] 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. 3, 4
- [17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2, 3
- [18] Levon Khachatryan, Andranik Mojsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15954–15964, 2023. 2, 3
- [19] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 7
- [20] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 2
- [21] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with llm. *arXiv preprint arXiv:2401.01256*, 2024. 3
- [22] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 2, 6

- [23] Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, Hyeokmin Kwon, and Sang-pil Kim. Mtvg: Multi-text video generation with text-to-video models. *arXiv preprint arXiv:2312.04086*, 2023. 2, 3
- [24] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 9, 10
- [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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 6, 7
- [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. 2, 4
- [27] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhui Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024. 3
- [28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 4
- [29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4
- [30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 16
- [31] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 7
- [32] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3, 4
- [33] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3
- [34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2
- [35] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5
- [37] 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 descriptions. In *International Conference on Learning Representations*, 2022. 2
- [38] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. 3
- [39] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2, 4, 5, 16
- [40] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 6
- [41] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 3
- [42] Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, Chong Luo, Yueyi Zhang, and Zhiwei Xiong. Art•v: Auto-regressive text-to-video generation with diffusion models. *arXiv preprint arXiv:2311.18834*, 2023. 3
- [43] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 2, 3, 9, 10, 15
- [44] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. *arXiv:2311.10982*, 2023. 3
- [45] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023. 2
- [46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 5, 6, 15
- [47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8
- [48] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qing, Xiang Wang, Deli Zhao, and

Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. 2023. 2, 3, 4, 6, 9,
10

Appendix

This appendix complements our main paper with further experiments, in which we investigate the text-to-video generation quality of StreamingT2V for even longer sequences than those assessed in the main paper, and it contains additional information on the implementation of StreamingT2V and the experiments carried out.

In Sec. 7, a user study is conducted on the test set, in which all text-to-video methods under consideration are evaluated by humans to determine the user preferences.

Sec. 8 supplements our main paper by additional qualitative results of StreamingT2V, qualitatively comparisons of StreamingT2V with competing methods, and ablation studies to show the effectiveness of APM and randomized blending.

In Sec. 9, hyperparameters used in StreamingT2V, and implementation details of our ablated models are provided

Finally, sec. 10 complements our main paper by additional information on the employed metrics used in our qualitative experiments and it provides the prompts that compose our testset.

7. User Study

We conduct a user study comparing our StreamingT2V method with prior work using the 50-prompt test set results obtained for Sec. 5.4 of the main paper. To remove potential biases, we resize and crop all videos to align them. The user study is structured as a one vs one comparison between our StreamingT2V method and competitors where participants are asked to answer three questions for each pair of videos:

- Which model has better motion?
- Which model has better text alignment?
- Which model has better overall quality?

We accept exactly one of the following three answers for each question: preference for the left model, preference for the right model, or results are considered equal. To ensure fairness, we randomize the order of the videos presented in each comparison, and the sequence of comparisons. Fig. 8 shows the preference score obtained from the user study as the percentage of votes devoted to the respective answer.

Across all comparisons to competing methods, StreamingT2V is significantly more often preferred than the competing method, which demonstrates that StreamingT2V clearly improves upon state-of-the-art for long video generation. For instance in motion quality, as the results of StreamingT2V are non-stagnating videos, temporal consistent and possess seamless transitions between chunks, 65% of the votes were preferring StreamingT2V, compared to 17% of the votes preferring SEINE.

Competing methods are much more affected by quality degradation over time, which is reflected in the preference for StreamingT2V in terms of *text alignment* and *overall*

quality.

8. Very Long Video Generation

Supplementing our main paper, we study the generation quality of our text-to-video method StreamingT2V and the competing methods for creating even longer videos. To this end, we generate 240 frames, and even more, 600 frame videos.

Within this setting, we present qualitative results of StreamingT2V, qualitative and quantitative comparisons to existing methods, and ablation studies on our APM module and the randomize blending approach.

8.1. Qualitative Results.

Fig. 10 and Fig. 11 show text-to-video results of StreamingT2V for different actions, e.g. *dancing*, *reading*, *camera moving*, or *eating*, and different characters like *tiger* or *Santa Claus*. We can observe that scene and object features are kept across each video generation (see e.g. Fig. 10 (e) and Fig. 11 (d)), thanks to our proposed APM module. Our proposed CAM module ensures that generated videos are temporally smooth, with seamless transitions between video chunks, and not stagnating (see e.g. Fig. 10 (f)).

8.2. Comparison to state-of-the-art.

We perform a comprehensive evaluation of all text-to-video methods (which we considered in the main paper) on the test set, generating 240 frame-long videos for each.

To further analyse the suitability of text-to-video methods for long video generation, we extend our test set evaluation of Sec. 5.4 of our main paper, where we were using 80 frames, to 240 frames. Fig. 9 shows the results evaluated from 80 frame-long sequences up to the 240 frames videos with a step size of 10 frames.

In accordance with our visual observations, we see that most competing methods suffer from heavily deteriorated text-image alignment (CLIP score) as they generate more frames. For instance, the CLIP score for SVD drops by more than 16%. But also DynamiCrafter-XL, SparseControl and I2VGen-XL cannot maintain image quality. We conjecture that this behaviour is to a high degree caused by the CLIP image encoder (which is used by SVD, DynamiCrafter-XL and I2VGen-XL). The CLIP image encoder has been trained on clean and natural images, but the conditioning is performed on generated frames, which might confuse the CLIP image encoder, leading eventually to severe quality degradation. In contrast to that, the APM module of StreamingT2V conditions on a fixed anchor frame, so that it does not suffer from error accumulation. Overall, StreamingT2V has the highest CLIP score among all evaluated methods throughout all 240 frames, thanks to an effective conditioning on the anchor frame using our APM module. The second best method in terms

on CLIP score is FreeNoise, which achieves a similar but slightly worse CLIP score. However, our shown qualitative results have revealed that FreeNoise is heavily prone to video stagnation, so that the quality is not getting worse, but the video content is also nearly not changed.

Fig. 9 also shows that StreamingT2V seamlessly connects video chunks and creates temporal consistent videos, thanks to our effective CAM module, that uses the temporal attentional mechanism to condition a new video generation on the previous chunk. The MAWE score of StreamingT2V is hardly influenced by longer video generation, thus leading to the best MAWE score among all considered methods, which is more than 50% better than the best competing method SEINE. The experiment also shows that the competing methods SVD, I2VGen-XL, SparseControl and FreeNoise are unstable. With more frames being generated, their MAWE score becomes worse. These methods are thus prone to increased temporal inconsistencies or video stagnation. This again shows that their conditioning mechanism is too weak (e.g., the concatenation of the conditional frame to the input and to the cross attention layers for SVD).

Finally, the SCuts score shows that StreamingT2V leads to the least scene cuts and thus best transitions. The advantage over SparseControl, DynamiCrafter-XL and SVD is significant, especially with more frames (e.g. compared to StreamingT2V the MAWE score is up to 77 times lower for StreamingT2V).

Our study on long video generation has thus shown that thanks to our CAM and APM module, StreamingT2V is best suited for long video generation. All other competing methods heavily deteriorate their video quality when applied for long video generation.

8.3. Ablation Studies

We present additional qualitative results for the ablations on our APM module and randomized blending.

Effectiveness of APM. We complement our ablation study on APM (see Sec. 5.3 of the main paper) by additional qualitative results in Fig. 12. Thanks to the usage of long-term information via our proposed APM module, identity and scene features are preserved throughout the video. For instance, the face of the woman (including all its tiny details) are consistent across the video generation. Also, the style of the jacket and the bag are correctly generated throughout the video. Without having access to a long-term memory, these object and scene features are changing over time.

Randomized Blending. Fig. 13 and Fig. 14 show additional ablated results on our randomized blending approach. From the X-T slice visualizations we can see that the randomized blending leads to smooth chunk transitions. In contrast, when naively concatenating enhanced video chunks, or using shared noise, the resulting videos possess visible inconsistencies between chunks.

9. Implementation details

We provide additional implementation details for StreamingT2V and our ablated models.

9.1. Streaming T2V Stage

For the StreamingT2V stage, we use classifier free guidance [10, 14] from text and the anchor frame. More precisely, let $\epsilon_\theta(x_t, t, \tau, a)$ denote the noise prediction in the StreamingT2V stage for latent code x_t at diffusion step t , text τ and anchor frame a . For text guidance and guidance by the anchor frame, we introduce weights ω_{text} and ω_{anchor} , respectively. Let τ_{null} and a_{null} denote the empty string, and the image with all pixel values set to zero, respectively. Then, we obtain the multi-conditioned classifier-free-guided noise prediction $\hat{\epsilon}_\theta$ (similar to DynamiCrafter-XL [43]) from the noise predictor ϵ via

$$\begin{aligned} \hat{\epsilon}_\theta(x_t, t, \tau, a) = & \epsilon_\theta(x_t, t, \tau_{\text{null}}, a_{\text{null}}) \\ & + \omega_{\text{text}} (\epsilon_\theta(x_t, t, \tau, a_{\text{null}}) \\ & - \epsilon_\theta(x_t, t, \tau_{\text{null}}, a_{\text{null}})) \\ & + \omega_{\text{anchor}} (\epsilon_\theta(x_t, t, \tau, a) \\ & - \epsilon_\theta(x_t, t, \tau, a_{\text{null}})). \end{aligned} \quad (9)$$

We then use $\hat{\epsilon}_\theta$ for denoising. In our experiments, we set $\omega_{\text{text}} = \omega_{\text{anchor}} = 7.5$. During training, we randomly replace τ with τ_{null} with 5% likelihood, the anchor frame a with a_{null} with 5% likelihood, and we replace at the same time τ with τ_{null} and a with a_{null} with 5% likelihood.

Additional hyperparameters for the architecture, training and inference of the Streaming T2V stage are presented in Tab. 9, where *Per-Pixel Temporal Attention* refers to the attention module used in CAM (see Fig. 2 of the main paper).

9.2. Ablation models

In Sec. 5.3 of the main paper, we consider two baselines that we compare with CAM. Here we provide additional implementation details.

The ablated model *Add-Cond* applies to the features of CAM (i.e. the outputs of the encoder and middle layer of the ControlNet part in Fig. 2) zero-convolution, and uses addition to fuse it with the features of the skip-connection of the UNet (similar to ControlNet[46]) (see Fig. 15). We provide here additional details to construct this model. Given a video sample $\mathcal{V} \in \mathbb{R}^{F \times H \times W \times 3}$ with $F = 16$ frames, we construct a mask $M \in \{0, 1\}^{F \times H \times W \times 3}$ that indicates which frame we use for conditioning, i.e. $M^f[i, j, k] = M^f[i', j', k']$ for all frames $f = 1, \dots, F$ and for all i, j, k, i', j', k' . We require that exactly $F - F_{\text{cond}}$ frames are masked, i.e.

$$\sum_{f=1}^F M^f[i, j, k] = F - F_{\text{cond}}, \text{ for all } i, j, k. \quad (10)$$

Per-Pixel Temporal Attention	
Sequence length Q	16
Sequence length K,V	8
Token dimensions	320,640,1280
Appearance Preservation Module	
CLIP Image Embedding Dim	1024
CLIP Image Embedding Tokens	1
MLP hidden layers	1
MLP inner dim	1280
MLP output tokens	16
MLP output dim	1024
1D Conv input tokens	93
1D Conv output tokens	77
1D Conv output dim	1024
Cross attention sequence length	77
Training	
Parametrization	ϵ
Diffusion Setup	
Diffusion steps	1000
Noise scheduler	Linear
β_0	0.0085
β_T	0.0120
Sampling Parameters	
Sampler	DDIM
Steps	50
η	1.0
ω_{text}	7.5
ω_{anchor}	7.5

Table 9. Hyperparameters of Streaming T2V Stage. Additional architectural hyperparameters are provided by the Modelsope report[39].

We concatenate $[\mathcal{V} \odot M, M]$ along the channel dimension and use it as input for the image encoder $\mathcal{E}_{\text{cond}}$, where \odot denotes element-wise multiplication.

During training, we randomly set the mask M . During inference, we set the mask for the first 8 frames to zero, and for the last 8 frames to one, so that the model conditions on the last 8 frames of the previous chunk.

For the ablated model *Conc-Cond* used in Sec. 5.3, we start from our Video-LDM’s UNet, and modify its first convolution. Like for *Add-Cond*, we consider a video \mathcal{V} of length $F = 16$ and a mask M that encodes which frames are overwritten by zeros. Now the Unet takes $[z_t, \mathcal{E}(\mathcal{V}) \odot M, M]$ as input, where we concatenate along the channel dimension. As with *Add-Cond*, we randomly set M during training so that the information of 8 frames is used, while during inference, we set it such that the last 8 frames of the previous chunk are used. Here \mathcal{E} denotes the VQ-GAN encoder (see Sec. 3.1 of the main paper).

10. Experiment details

10.1. Metric details

In Sec. 5.3 of the main paper, we employ a re-identification score to assess the feature preservation quality of our APM module in the *Long-term memory* experiment.

To obtain the score, let $P_n = \{p_i^n\}$ be all face patches extracted from frame n using an off-the-shelf head detector [30] and let F_i^n be the corresponding facial feature of p_i^n , which we obtain from an off-the-shelf face recognition network [30]. Then, for frame n , $n_1 := |P_n|$, $n_2 := |P_{n+1}|$, we define the re-id score $\text{re-id}(n)$ for frame n as

$$\text{re-id}(n) := \begin{cases} \max_{i,j} \cos \Theta(F_i^n, F_j^{n+1}), & n_1 > 0 \& n_2 > 0. \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

where $\cos \Theta$ is the cosine similarity. Finally, we obtain the re-ID score of a video by averaging over all frames, where the two consecutive frames have face detections, i.e. with $m := |\{n \in \{1, \dots, N\} : |P_n| > 0\}|$, we compute the weighted sum:

$$\text{re-id} := \frac{1}{m} \sum_{n=1}^{N-1} \text{re-id}(n), \quad (12)$$

where N denotes the number of frames.

MAWE is based on Warp errors and OF scores which have highly dependent values. We try to counteract this effect in our MAWE score by assuming this dependency is linear $W(\mathcal{V}) = c \cdot \text{OFS}(\mathcal{V})$ and account for it in MAWE’s denominator. To calculate c we randomly sample a small part of our training with a range of optical scores and remove outliers applying the Z-score method. Using this dataset c is obtained by a simple regression analysis.

10.2. Test prompt set

1. A camel resting on the snow field.
2. Camera following a pack of crows flying in the sky.
3. A knight riding on a horse through the countryside.
4. A gorilla eats a banana in Central Park.
5. Men walking in the rain.
6. Ants, beetles and centipede nest.
7. A squirrel on a table full of big nuts.
8. Close flyover over a large wheat field in the early morning sunlight.
9. A squirrel watches with sweet eyes into the camera.
10. Santa Claus is dancing.
11. Chemical reaction.
12. Camera moving in a wide bright ice cave, cyan.
13. Prague, Czech Republic. Heavy rain on the street.
14. Time-lapse of stormclouds during thunderstorm.
15. People dancing in room filled with fog and colorful lights.

16. Big celebration with fireworks.
17. Aerial view of a large city.
18. Wide shot of battlefield, stormtroopers running at night, fires and smokes and explosions in background.
19. Explosion.
20. Drone flythrough of a tropical jungle with many birds.
21. A camel running on the snow field.
22. Fishes swimming in ocean camera moving.
23. A squirrel in Antarctica, on a pile of hazelnuts cinematic.
24. Fluids mixing and changing colors, closeup.
25. A horse eating grass on a lawn.
26. The fire in the car is extinguished by heavy rain.
27. Camera is zooming out and the baby starts to cry.
28. Flying through nebulas and stars.
29. A kitten resting on a ball of wool.
30. A musk ox grazing on beautiful wildflowers.
31. A hummingbird flutters among colorful flowers, its wings beating rapidly.
32. A knight riding a horse, pointing with his lance to the sky.
33. steampunk robot looking at the camera.
34. Drone fly to a mansion in a tropical forest.
35. Top-down footage of a dirt road in forest.
36. Camera moving closely over beautiful roses blooming time-lapse.
37. A tiger eating raw meat on the street.
38. A beagle looking in the Louvre at a painting.
39. A beagle reading a paper.
40. A panda playing guitar on Times Square.
41. A young girl making selfies with her phone in a crowded street.
42. Aerial: flying above a breathtaking limestone structure on a serene and exotic island.
43. Aerial: Hovering above a picturesque mountain range on a peaceful and idyllic island getaway.
44. A time-lapse sequence illustrating the stages of growth in a flourishing field of corn.
45. Documenting the growth cycle of vibrant lavender flowers in a mesmerizing time-lapse.
46. Around the lively streets of Corso Como, a fearless urban rabbit hopped playfully, seemingly unfazed by the fashionable surroundings.
47. Beside the Duomo's majestic spires, a fearless falcon soared, riding the currents of air above the iconic cathedral.
48. A graceful heron stood poised near the reflecting pools of the Duomo, adding a touch of tranquility to the vibrant surroundings.
49. A woman with a camera in hand joyfully skipped along the perimeter of the Duomo, capturing the essence of the moment.
50. Beside the ancient amphitheater of Taormina, a group of friends enjoyed a leisurely picnic, taking in the breath-
- taking views.

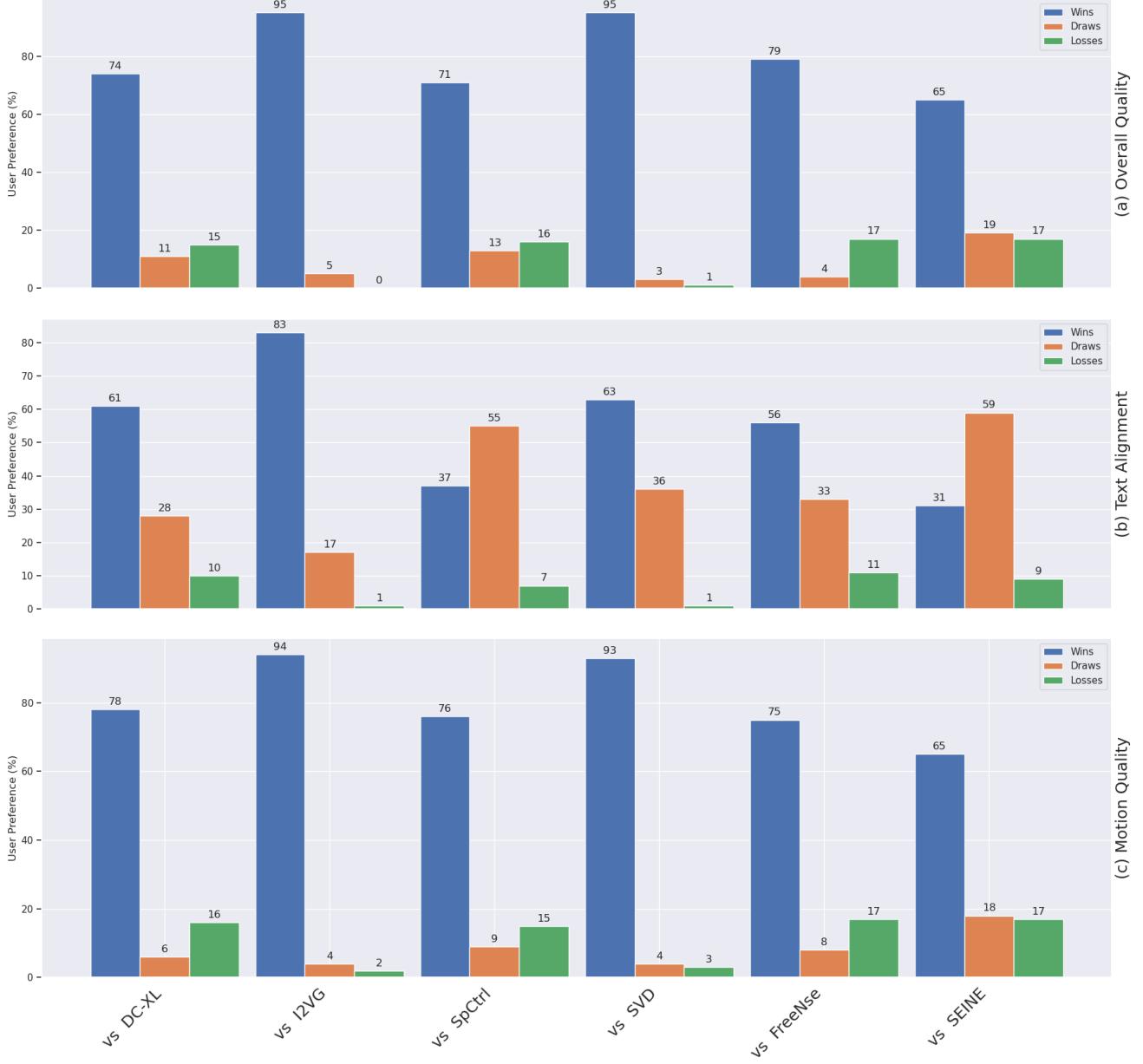


Figure 8. We conduct a user study, asking humans to assess the test set results of Sect 5.4 in a one-to-one evaluation, where for any prompt of the test set and any competing method, the results of the competing method have to be compared with the corresponding results of our StreamingT2V method. For each comparison of our method to a competing method, we report the relative of number votes that prefer StreamingT2V (i.e. wins), that prefer the competing method (i.e. losses), and that consider results from both methods as equal (i.e. draws).

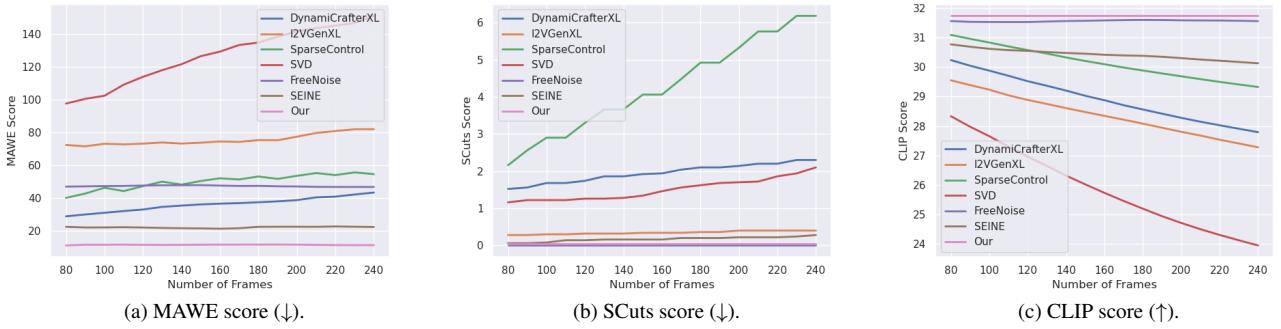


Figure 9. Study on how generating longer videos are affecting the generation quality.

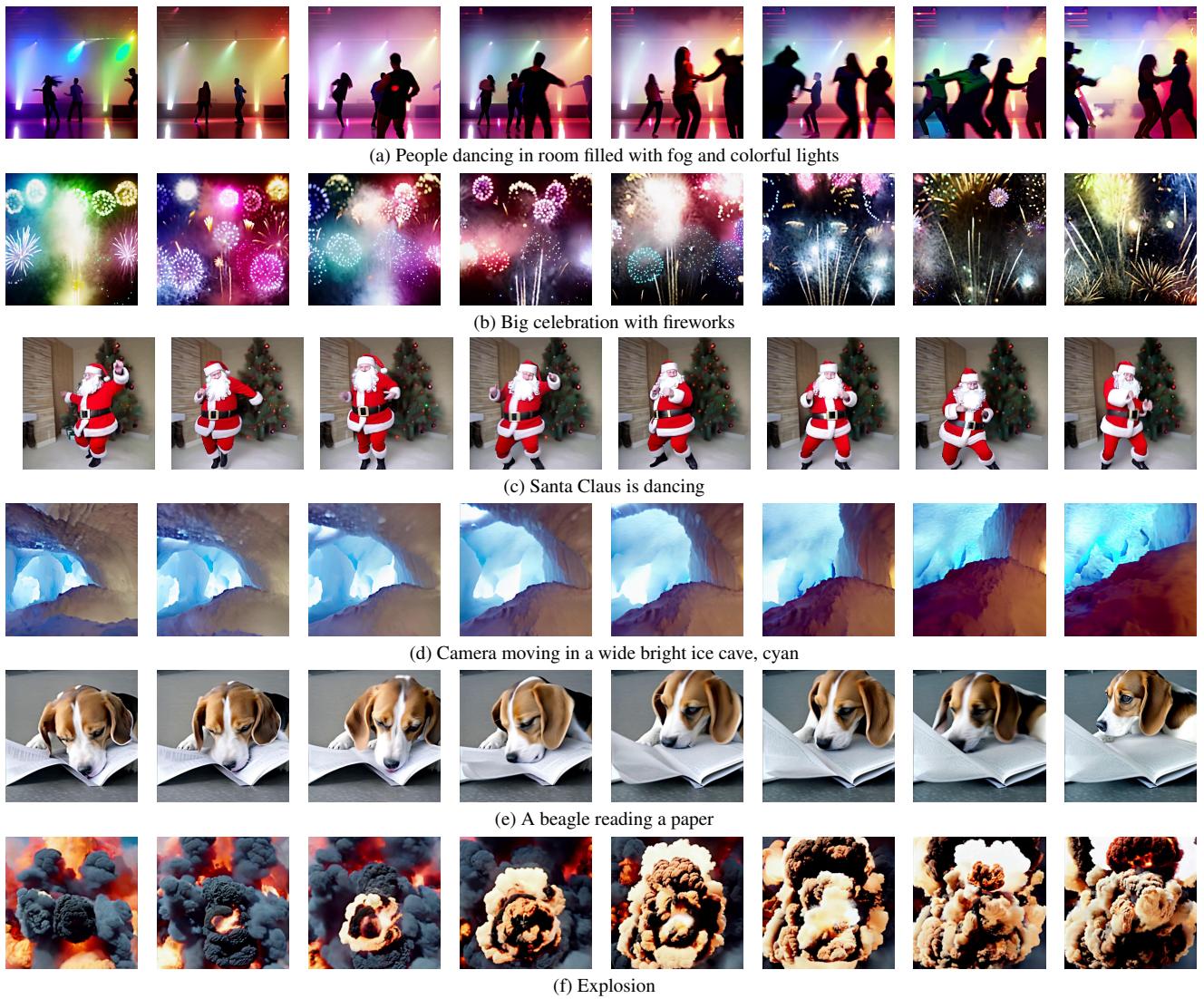


Figure 10. Qualitative results of StreamingT2V for different prompts.

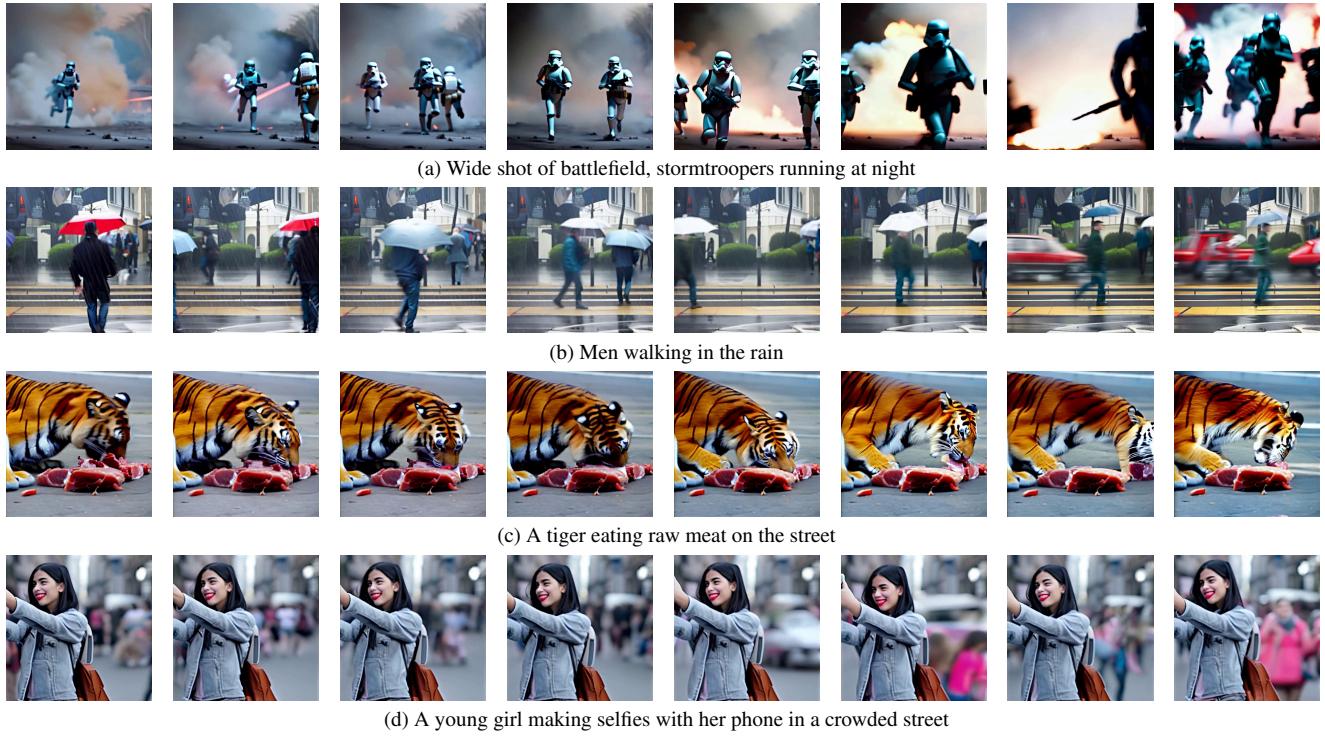


Figure 11. Qualitative results of StreamingT2V for different prompts.

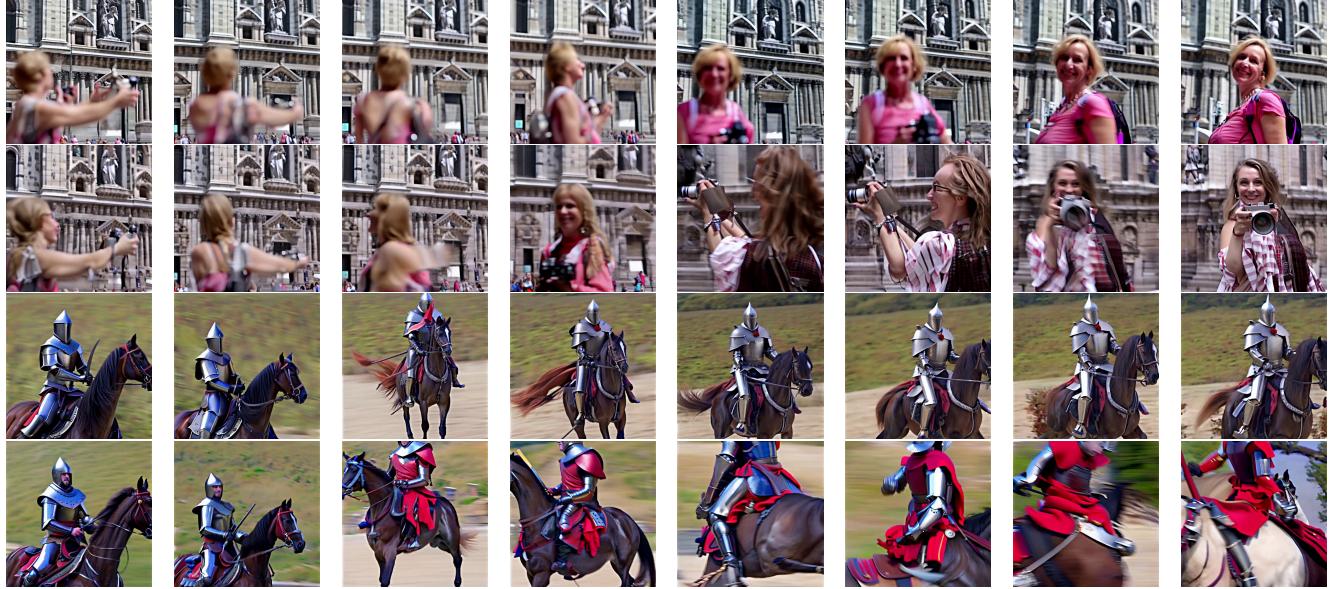


Figure 12. Ablation study on the APM module. Top row is generated from StreamingT2V, bottom row is generated from StreamingT2V w/o APM.

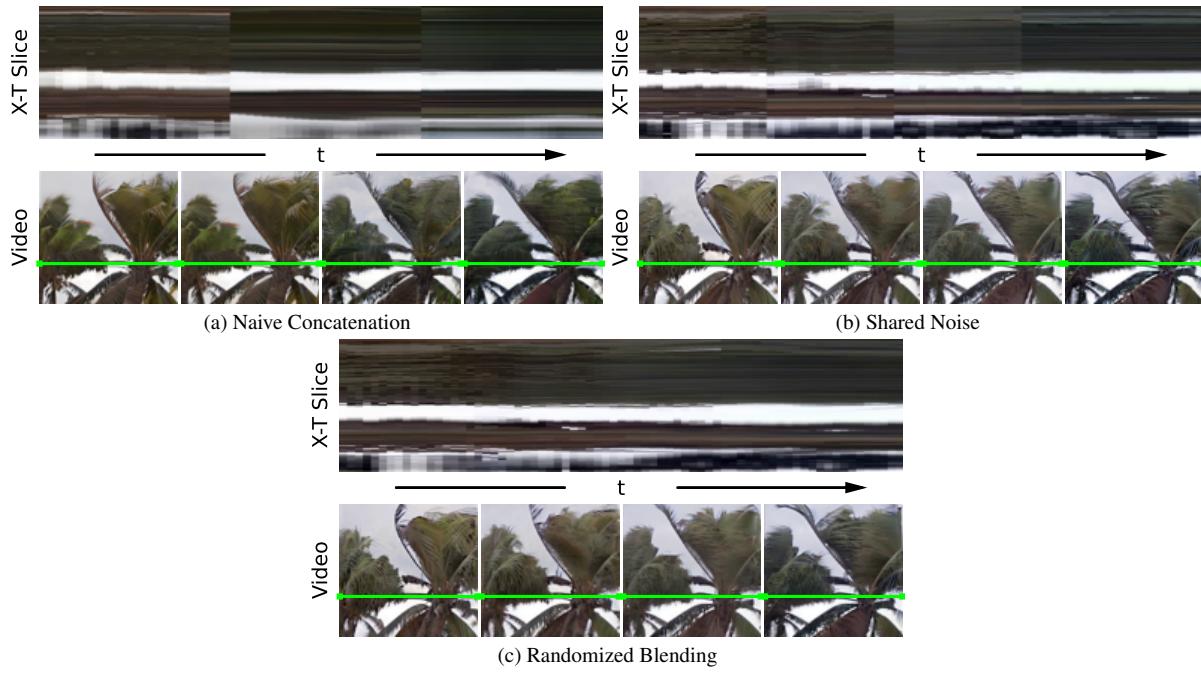


Figure 13. Ablation study on different approaches for autoregressive video enhancement.

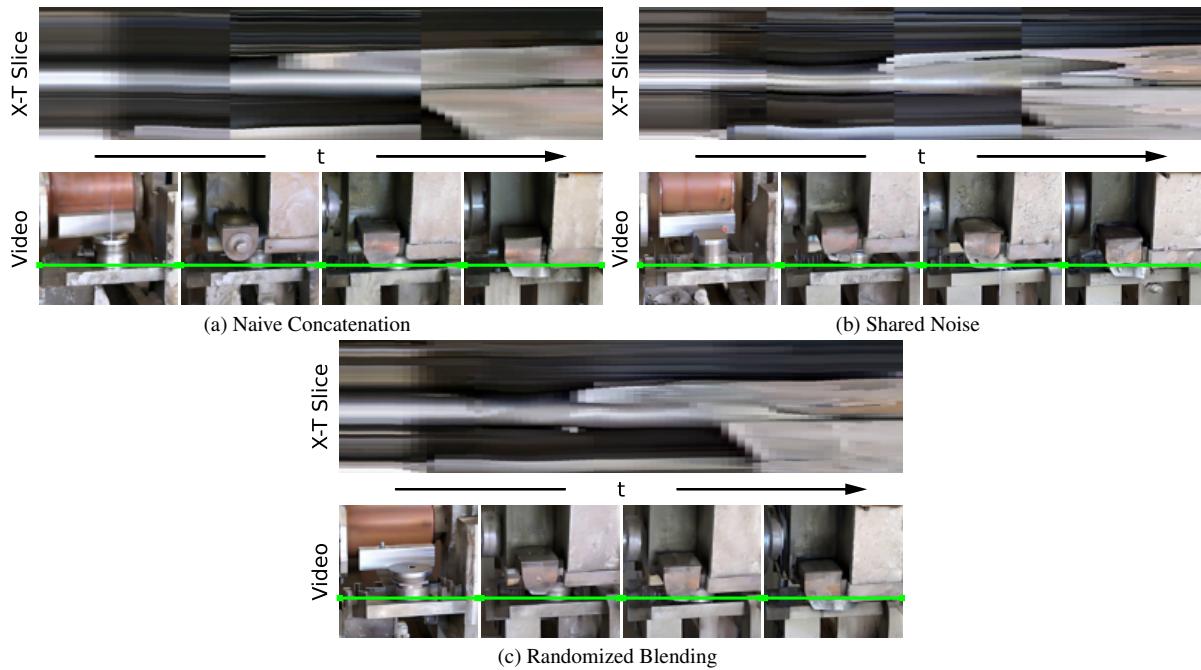


Figure 14. Ablation study on different approaches for autoregressive video enhancement.

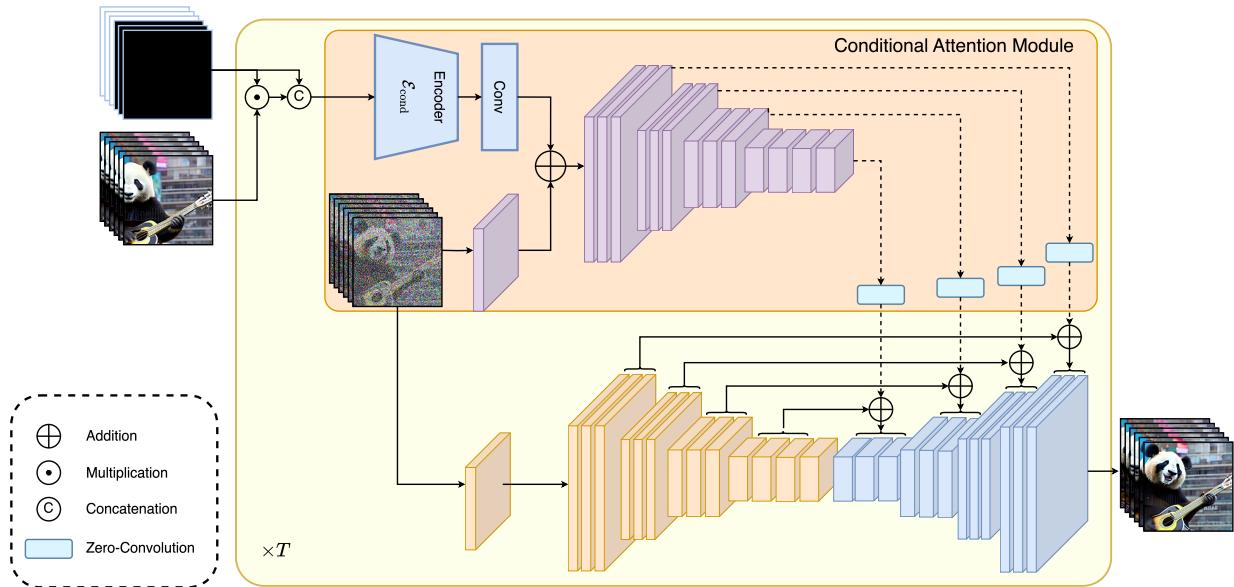


Figure 15. Illustration of the Add-Cond baseline, which is used in Sec. 5.3 of the main paper.