
LTX-Video: Realtime Video Latent Diffusion

Yoav HaCohen	Nisan Chiprut	Benny Brazowski	Daniel Shalem
Dudu Moshe	Eitan Richardson	Eran Levin	Guy Shiran
Nir Zabari	Ori Gordon	Poriya Panet	Sapir Weissbuch
Victor Kulikov	Yaki Bitterman	Zeev Melumian	Ofir Bibi*

Lightricks
ltx-video@lightricks.com

Abstract

We introduce *LTX-Video*, a transformer-based latent diffusion model that adopts a holistic approach to video generation by seamlessly integrating the responsibilities of the Video-VAE and the denoising transformer. Unlike existing methods, which treat these components as independent, *LTX-Video* aims to optimize their interaction for improved efficiency and quality. At its core is a carefully designed Video-VAE that achieves a high compression ratio of 1:192, with spatiotemporal downscaling of $32 \times 32 \times 8$ pixels per token, enabled by relocating the patchifying operation from the transformer’s input to the VAE’s input. Operating in this highly compressed latent space enables the transformer to efficiently perform full spatiotemporal self-attention, which is essential for generating high-resolution videos with temporal consistency. However, the high compression inherently limits the representation of fine details. To address this, our VAE decoder is tasked with both latent-to-pixel conversion and the final denoising step, producing the clean result directly in pixel space. This approach preserves the ability to generate fine details without incurring the runtime cost of a separate upsampling module. Our model supports diverse use cases, including text-to-video and image-to-video generation, with both capabilities trained simultaneously. It achieves faster-than-real-time generation, producing 5 seconds of 24 fps video at 768x512 resolution in just 2 seconds on an Nvidia H100 GPU, outperforming all existing models of similar scale. The source code and pre-trained models are publicly available², setting a new benchmark for accessible and scalable video generation.

1 Introduction

The rise of text-to-video models such as [Sora](#) [1], [MovieGen](#) [2], [CogVideoX](#) [3], [Open-Sora Plan](#) [4] and [PyramidFlow](#) [5] has demonstrated the effectiveness of spatiotemporal transformers with self-attention and a global receptive field, coupled with 3D VAEs for spatiotemporal compression. While these approaches validate the fundamental architectural choices, they often rely on conventional VAE designs that may not optimally balance spatial and temporal compression.

Concurrently with our work, [DC-VAE](#) [6] demonstrated that text-to-image transformer-based diffusion models perform more effectively when paired with VAEs that employ higher spatial compression factors and a high dimensional latent spaces with up to 64 channels. However, extending this approach to video presents significant challenges.

^{*}Authors are listed with project leads first, followed by the team in alphabetical order, and concluding with senior management.

²<https://github.com/Lightricks/LTX-Video>.



"A young man with blonde hair wearing a yellow jacket stands in a forest and looks around. He has light skin and his hair is styled with a middle part. He looks to the left and then to the right, his gaze lingering in each direction. The camera angle is low, looking up at the man, and remains stationary throughout the video. The background is slightly out of focus, with green trees and the sun shining brightly behind the man. The lighting is natural and warm, with the sun creating a lens flare that moves across the man's face. The scene is captured in real-life footage."



Figure 1: Text-to-video (first row) and image-to-video samples (last 2 rows, conditioned on the left frame) generated by *LTX-Video*, highlighting our model’s high level of prompt adherence, visual quality and motion fidelity. Each row shows evenly-spaced frames from a generated 5-second video.

Inspired by these developments and the success of diffusion models in generating high-resolution images and video, we propose *LTX-Video*, a transformer-based latent diffusion model that equally prioritizes both spatial and temporal dimensions. Our approach features a carefully designed VAE architecture that achieves higher spatial compression while maintaining video quality through an increased latent depth of 128 channels. This design choice not only enables more efficient processing of video data but also results in a highly performant 3D VAE implementation.

Latent diffusion models trade the ability to apply pixel-level training loss for improved training efficiency, often at the expense of generating plausible high-frequency details. [Sora](#) [1] and [MovieGen](#) [2] mitigate this limitation by applying a second-stage diffusion model for generating the high-resolution output. [Pixel-loss](#) [7] attempt to address this issue by incorporating pixel-level loss on VAE-decoded noisy latents, but retained the entire generation process within the limits of the compressed latent space. In contrast, we propose tasking the VAE decoder with performing the *last denoising step* in conjunction with converting latents to pixels. This modification is particularly impactful at high latent compression rates, where not all high-frequency details can be reconstructed and must instead be generated.

We adopt the scalable and flexible transformer architecture, known for its effectiveness in various applications, enabling our model to generate images and videos across a range of sizes and durations. Building upon [Pixart- \$\alpha\$](#) [8]’s architecture, which extends the [DiT](#) [9] framework to be conditioned on open text inputs rather than constrained to ImageNet class labels, we introduce several key enhancements. Specifically, we replace traditional absolute positional embeddings with Rotary Positional Embeddings ([RoPE](#) [10]) enhanced by normalized fractional coordinates, which improve spatial and temporal coherence in video generation. Additionally, we normalize the *key* and *query* tensors to stabilize attention computations, enhancing robustness and increasing the entropy of attention weights. Our approach addresses limitations in existing models, offering a more integrated and efficient solution for robust video generation.

Our model is the fastest video generation model of its kind, capable of generating videos faster than the time it takes to watch them (2 seconds to generate 121 frames at 768×512 pixels and 20 diffusion steps on an Nvidia H100 GPU), while outperforming all available models of similar scale (2B parameters, before distillation).

In addition to text-to-video generation, we extend our model’s functionality to handle image-to-video, a practical application in content creation. Through a simple timestep-based conditioning mechanism, the model can be conditioned on any part of the input video without requiring additional parameters or special tokens.

See Fig 1 for text-to-video and image-to-video samples generated by *LTX-Video*. Additional samples are provided in figures 18 and 19.

Our main contributions are:

- **A holistic approach to latent diffusion:** *LTX-Video* seamlessly integrates the Video-VAE and the denoising transformer, optimizing their interaction within a compressed latent space and sharing the denoising objective between the transformer and the VAE’s decoder.
- **High-compression Video-VAE leveraging novel loss functions:** By relocating the patchifying operation to the VAE and introducing novel loss functions, we achieve a 1:192 compression ratio with spatiotemporal downsampling of $32 \times 32 \times 8$, enabling the generation of high-quality videos at unprecedented speed.
- ***LTX-Video* – A fast, accessible, and high-quality video generation model:** We train and evaluate our enhanced diffusion-transformer architecture and publicly release *LTX-Video*, a faster-than-real-time text-to-video and image-to-video model with fewer than 2B parameters.

2 Method

To facilitate the faster-than-realtime operation of *LTX-Video* while maintaining high visual quality, motion fidelity, and prompt adherence, we employ a holistic approach to latent diffusion, optimizing the interaction between the Video VAE and the diffusion-transformer. We utilize a high-dimensional latent space with a high compression rate of 1:192 and spatiotemporal downsampling of $32 \times 32 \times 8$. To support the generation of high-frequency details, we assign the VAE decoder the task of performing the last denoising step alongside converting the latents to pixels, as illustrated in Fig 2.

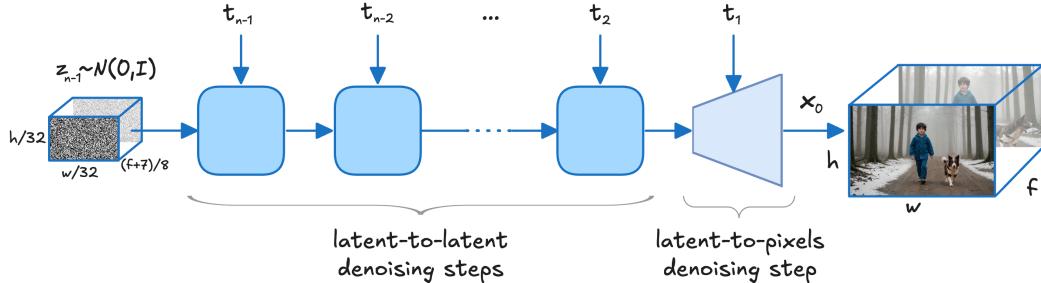


Figure 2: *LTX-Video* holistic denoising strategy – latent-to-latent denoising steps + final latent-to-pixels denoising step.

Our holistic approach, along with key design changes in the Video VAE architecture, loss functions, and in the diffusion-transformer architecture, enable generating high quality videos despite the high pixels-to-tokens ratio. These enhancements, which were crucial to the success of our approach, are highlighted in the following sections.

2.1 Video VAE

Operating in a compressed latent space is key to the success of text-to-video diffusion-transformer models: training and inference time for these models is dominated by the number of tokens (the attention operation is quadratic in the number of tokens), while the diffusion process benefits from the compressed latent representation, as it decreases the inherent information redundancy of the original signal ([SimpleDiffusion](#) [11], [SD3](#) [12]).

Recent text-to-video models ([CogVideoX](#) [3], [MovieGen](#) [2], [PyramidFlow](#) [5], [Open-Sora Plan](#) [4], [HunyuanVideo](#) [13]) employ VAEs that downscale the spatio-temporal dimensions by either $8 \times 8 \times 4$

or $8 \times 8 \times 8$, while increasing the number of channels from 3 to 16. These configurations result in a total compression factor of 1:48 or 1:96. Subsequently, a *patchifier* collects latent patches of size $2 \times 2 \times 1$ into tokens, achieving an effective pixels-to-tokens ratio of 1:1024 or 1:2048, at the input of the transformer.

In contrast, our Video-VAE applies a spatio-temporal compression of $32 \times 32 \times 8$ with 128 channels, resulting in a total compression of 1:192 (twice the typical compression) and a pixels-to-tokens ratio of 1:8192 (four times the typical ratio), without requiring a patchifier. See Table 1 for additional details.

The challenge of information redundancy in pixel space at high resolutions has been highlighted by **SimpleDiffusion** [11], who mitigated it by increasing the amount of noise added at each diffusion step. This challenge also applies to higher-resolution latents (**SD3** [12]), as well as to redundancy in the time dimension. Therefore, spatial and temporal compression is crucial.

We analyze the redundancy in our latent space using **PCA** [14] over the latent pixels of 128 video samples (see Fig 3). As training progresses, our VAE learns to utilize the available channels and reduces their redundancy. Note that naive patchification of the latents before passing them to the transformer, as done by recent models, does not contribute to reducing the redundancy.

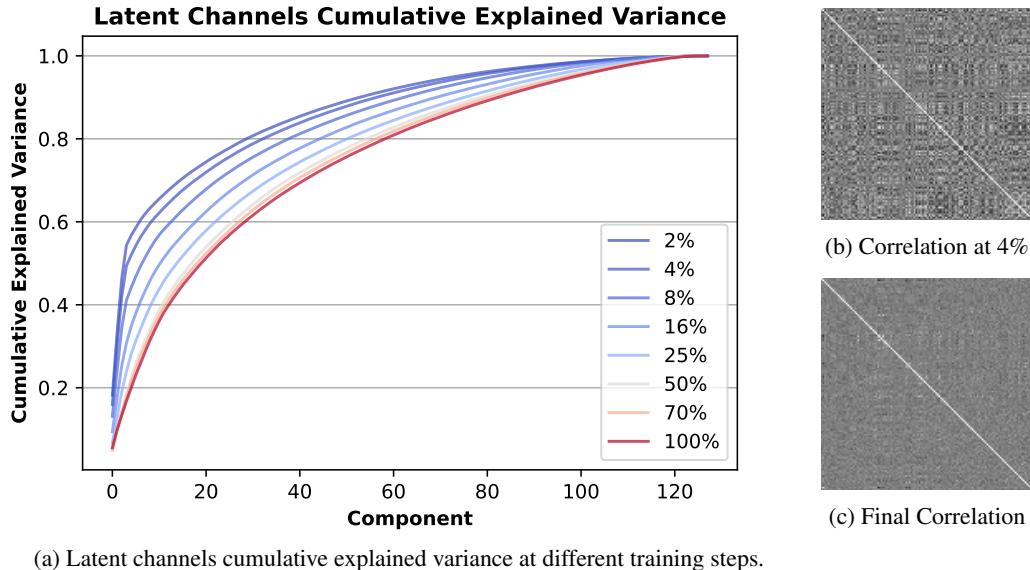


Figure 3: Latent-space redundancy. (a) Cumulative explained-variance of latent channels at different training steps (2% - 100% of training). As training progresses, the redundancy reduces and components contribute more evenly to the variance. (b, c) Latent channels auto-correlation matrices: high off-diagonal values early (at 4% of total training steps) and near-zero at training completion.

This pixels-to-latents compression is a critical process in our approach and the main enabler of its unprecedented speed.

To facilitate the high-compression rate without loss of quality, we introduce several key enhancements over current VAEs, as described in the following sections. We trained and compared several VAEs designed to efficiently map both videos and images into a unified latent space. These VAEs represent a key component of our model. See Fig 4 for our VAE architecture. Notably, compared to standard DiT diffusion models, we move the patchifying layer from the beginning of the transformer to the beginning of the VAE encoder and task the VAE decoder with performing the last denoising step in conjunction with decoding the latents into pixels.

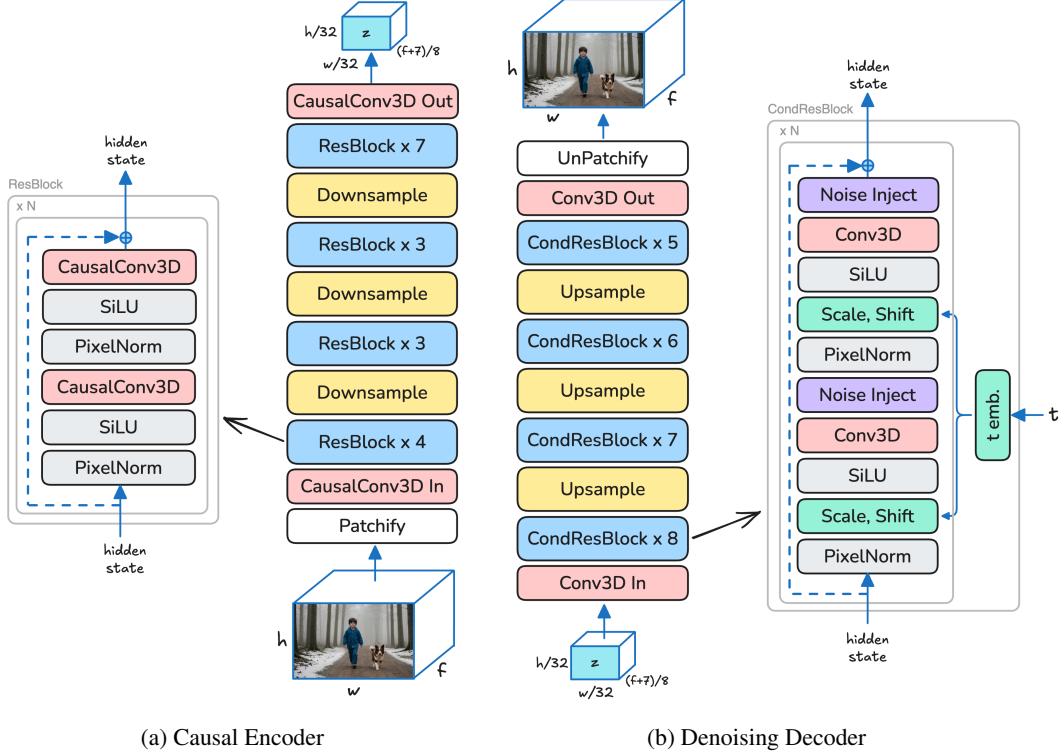


Figure 4: The *LTX-Video* Video-VAE architecture: (a) *Causal Encoder* utilizing 3D Causal Convolutions, applying $32 \times 32 \times 8$ compression (except the first frame, which is encoded as a separate latent frame). (b) *Denoising Decoder* with diffusion-timestep conditioning and multi-layer noise injection.

2.1.1 Shared Diffusion Objective

Rectified-flow models are designed to map noisy latents, $z_{t_i} = (1 - t_i)z_0 + t_i\epsilon$, to expected clean latents, z_0 , such that $f^\theta(z_{t_i}, t_i) = z_0$ ³. By initializing z_0 as pure noise (ϵ) and iteratively updating z_0 , with decreasing noise levels, $\{t_N, t_{N-1}, t_{N-2}, \dots, t_2, t_1\}$, the predicted z_0 becomes progressively cleaner until it closely matches the distribution of the training set.

In practice, however, the number of iterations is limited, and true convergence is rarely achieved. Consequently, residual uncertainty persists.

In *latent* diffusion, where $z_0 = \mathcal{E}(x_0)$ is a compressed representation of the data sample x_0 , this residual uncertainty often manifests as out-of-distribution inputs to the decoder \mathcal{D} , resulting in artifacts in the reconstructed pixel space, $x_0 = \mathcal{D}(z_0)$. Our experiments show that this effect is exacerbated when the encoder \mathcal{E} operates at high compression rates, particularly in regions with high-frequency signals that are poorly represented in the compressed latent space.

As operating within a compressed latent space is critical for the practical usability of video diffusion transformers, various strategies have been proposed to address these artifacts. For instance, **MovieGen** [2] introduced a diffusion-based upsampler operating in a less aggressively compressed latent space, while **Sora** [1] proposed an upsampler working directly in pixel space, conditioned on the latent outputs of the base model. While these approaches effectively mitigate artifacts, they incur significant computational and runtime costs.

To generate precise fine details while retaining faster generation times, we propose a novel approach that fuses the decoding and denoising steps. Specifically, we train the decoder as a diffusion model that maps noisy latents to clean pixels at varying noise levels: $x_0 = \mathcal{D}(z_{t_i}, t_i) = \mathcal{D}((1 - t_i)z_0 + t_i\epsilon, t_i)$. Since \mathcal{D} maps between spaces of different dimensionality, it cannot be applied iteratively like a

³In most implementations, the model f^θ predicts an intermediate term from which z_0 can be inferred in closed form.

standard diffusion model. However, it can execute the final denoising step, $x_0 = \mathcal{D}(z_{t_1}, t_1)$, in a manner inaccessible to the base model. Unlike the latent-to-latent denoising base model, constrained by the limited expressiveness of the latent space, our denoising-decoder directly outputs in pixel space and is trained with pixel-space losses.

Our implementation of $\mathcal{D}(z_{t_i}, t_i)$ follows a standard coarse-to-fine latent-to-pixel decoding architecture. To condition on the timestep t_i , we employ adaptive normalization layers, as commonly used in U-Net-based diffusion models (e.g. [DDPM](#) [15]). The denoising-decoder is trained with noise levels in the range $[0, 0.2]$ corresponding to the final diffusion timestep in common noise schedulers.

2.1.2 Reconstruction GAN (rGAN)

A common approach in VAE training is to balance pixel-wise L2 loss, perceptual loss ([LPIPS](#) [16]), and a GAN discriminator. At high compression rates, L2 loss often produces blurry outputs. The addition of perceptual loss reduces blurriness but can introduce texture artifacts, particularly in high-motion scenarios. However, the adversarial training approach typically relies on discriminators designed for tasks unrelated to reconstruction. These discriminators are tasked with distinguishing between real and fake samples without additional context, making their job unnecessarily challenging for reconstruction-specific tasks. This challenge is particularly pronounced for Patch-GAN discriminators, which have a spatially limited context. For instance, it may be difficult for the discriminator to determine whether a blurry patch is due to depth-of-field effects or if it originates from a fake sample.

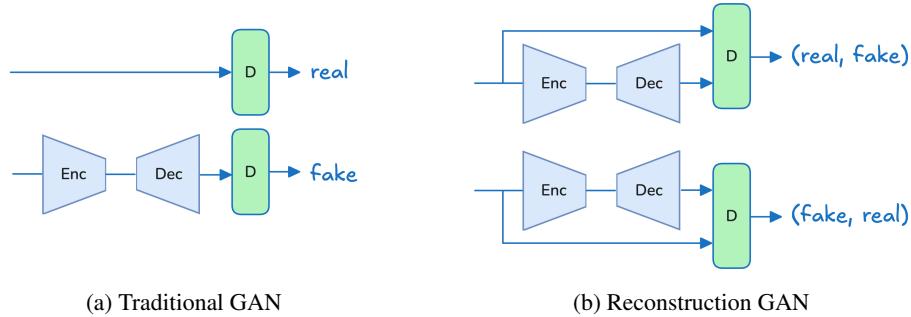


Figure 5: Our novel *Reconstruction GAN* loss. (a) Traditional GAN – the discriminator sees either a real or a reconstructed image. (b) Reconstruction GAN – the discriminator sees both versions **of the same sample** (concatenated) and needs to decide which is the original and which is the reconstructed version.

To address this, we propose *Reconstruction GAN* (see Fig 5), an adaptation of the traditional GAN training framework tailored to reconstruction tasks. In our approach, the discriminator is provided with both the input and the reconstructed samples for each iteration. Its goal is to determine which sample is the original (real) and which is reconstructed (fake). This relative comparison significantly simplifies the discriminator’s task and improves its ability to guide the generator.

Our experiments demonstrate that our proposed Reconstruction GAN greatly enhances GAN stability and performance. Moreover, it allows the discriminator to serve not only as a loss for matching the general distribution of real samples but also as a robust reconstruction loss, effectively balancing fidelity and perceptual quality.

2.1.3 Multi-layer Noise Injection

In current VAEs ([SD-XL](#) [17], [DC-VAE](#) [6]), stochasticity is introduced only by adding noise to the latents (according to the predicted log-variance values). Following [StyleGAN](#) [18], we also inject noise at several layers of the VAE decoder allowing the generation of more diverse high-frequency details. The noise level is learned per-channel.

2.1.4 Uniform log-variance

We found that when using a wide latent space (large number of channels), standard KL loss tends to result in uneven latent space where some of the channels are not utilized for reconstruction but are

“sacrificed” for satisfying the KL loss (the predicted means shrink and become close to zero and the predicted variances become close to one). To mitigate this issue we used a uniform variance for all latent channels – a single predicted *logvar* shared across channels. This uniformly distributes the effect of the KL loss across all channels.

2.1.5 Video DWT Loss

To mitigate the known insufficiency of the L1 or L2 pixel loss in ensuring the reconstruction of high-frequency details, we introduced a spatio-temporal Discrete Wavelet Transform (DWT) loss. Specifically, we compute 8 3D DWT transforms for both the input and reconstructed videos and use their L1 distance as the loss.

2.1.6 VAE Implementation Details

The final set of losses used for training our VAE is: pixel reconstruction (MSE), Video-DWT (L1), Perceptual (LPIPS) and Reconstruction-GAN.

We tested both causal and non-causal VAEs. While non-causal VAEs are easier to train for better reconstruction, causal VAEs enable easier simultaneous training on images and videos and first-frame conditioned video generation.

We also tested architectures that employ separable convolutions, utilizing 2D spatial kernels followed by 1D temporal kernels. We found the 3D convolutions to work slightly better.

2.2 Video Transformer

In an effort to optimize our transformer architecture for modeling diverse and complex data, we have incorporated several key modifications over the [Pixart- \$\alpha\$](#) [8] baseline, which align with recent advancements in the field. See Fig 6 for the architecture of our 3D transformer block.

2.2.1 Rotary Positional Embedding (RoPE)

Following the practices established by [FiT](#) [19], [LargeDiT](#) [20], and [H-DiT](#) [21] for image generation, we replaced traditional absolute positional embeddings with Rotary Positional Embeddings ([RoPE](#) [10]). RoPEs allow for a more dynamic and context-aware interpretation of positional information, which is crucial for managing the sequences of varying lengths and dimensions typical in video data.

In our positional embedding experiments, we tested three primary variants: (1) absolute positional embeddings, (2) RoPE with fractional coordinates, and (3) RoPE with fractional coordinates normalized by predefined maximum coordinates. See Fig 7. Our empirical results demonstrated superior performance with the normalized fractional coordinate approach (Fig 7c). To ensure consistency over different resolutions, number of frames and frame rate, we compute the spatial and temporal RoPE coordinates in pixels and seconds respectively, relative to predefined max resolution and duration. Incorporating the original FPS into the temporal embedding allows the model to generate more natural-looking motion.

While many open-source implementations use inverse exponential frequency spacing, our experiments, including controlled toy problems, revealed better performance with exponentially increasing frequencies (Fig 8). This finding aligns with recent theoretical work [22] suggesting that truncating lower frequencies can improve model performance.

2.2.2 QK Normalization

Following the findings presented in [ScalingViT](#) [23] and [LargeDiT](#) [20], we apply a normalization layer to the queries and keys before the dot product attention computation to avoid extremely large values in attention logits, which lead to attention weights with near-zero entropy.

We compared RMSNorm and [LayerNorm](#) [24] and found that RMSNorm performs better.

These enhancements are designed to significantly improve the model’s performance and adaptability, particularly in addressing the unique challenges presented by video and image generation tasks. By

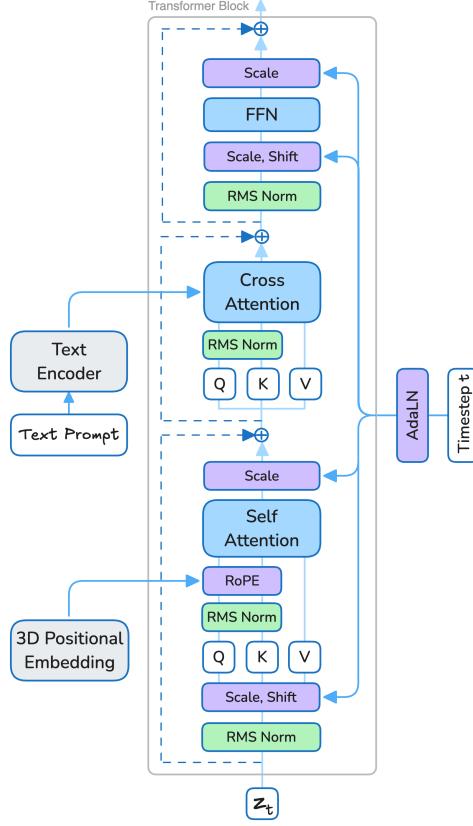


Figure 6: The *LTX-Video* 3D transformer-block architecture. Our architecture builds upon [Pixart- \$\alpha\$](#) [8], replacing LayerNorm with RMSNorm and incorporating QK-normalization and RoPE positional embeddings.

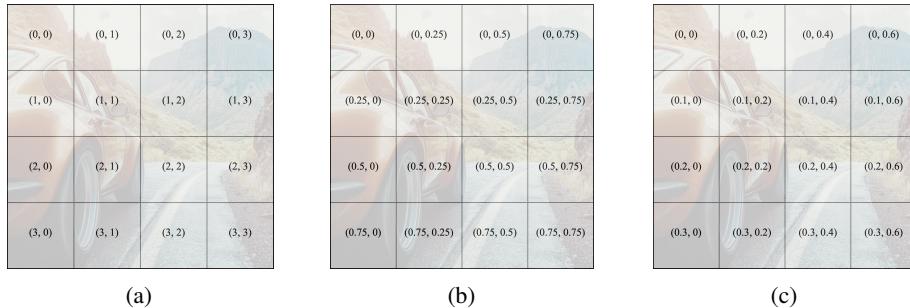


Figure 7: Positional encoding options: (a) Absolute positional encoding. (b) Fractional positional encoding. (c) Relative fractional positional encoding. Our experiments showed that relative-fractional positional embedding (option c) works best.

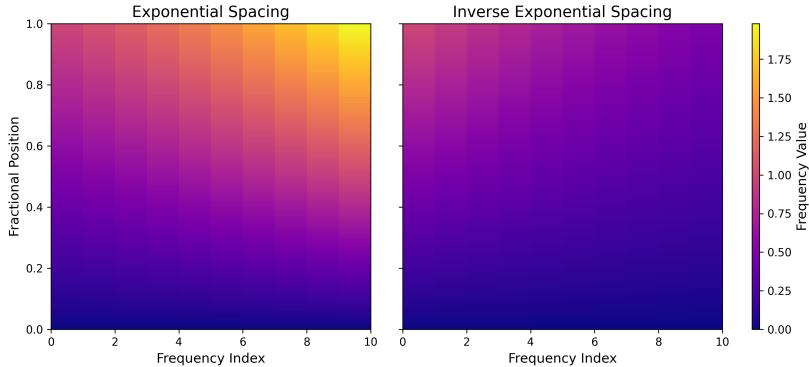


Figure 8: Different options for RoPE frequency spacing – exponential (left) and inverse-exponential (right). *LTX-Video* uses exponential spacing. See also section 4.3.2.

integrating these advanced features, our model achieves a higher degree of accuracy and efficiency in synthesizing high-quality visual content.

See Table 1 for a comparison of the *LTX-Video* architecture to recent text-to-video models - [MovieGen](#) [2], [HunyuanVideo](#) [13], [PyramidFlow](#) [5] and [CogVideoX](#) [3].

	MovieGen	HunyuanVideo	PyramidFlow	CogVideoX	LTX-Video
Base model size	30B	13B	2B	2B / 5B	1.9B
Transformer hidden dimension	6144	3072	1536	1920 / 3072	2048
FFN dimension factor	$2/3 \times 4$	4	4	4	4
VAE spatio-temporal compression	$8 \times 8 \times 8$	$8 \times 8 \times 4$	$8 \times 8 \times 8$	$8 \times 8 \times 4$	$32 \times 32 \times 8$
VAE output channels	16	16	16	16	128
VAE total compression	1:96	1:48	1:96	1:48	1:192
Transformer input patchifier	$2 \times 2 \times 1$	$1 \times 1 \times 1$			
Tokens to pixels ratio	1:2048	1:1024	1:2048	1:1024	1:8192
Number of Transformer blocks	48	60	24	30 / 42	28
Attention Blocks Architecture	Self + Cross	Self-only	Self-only	Self-only	Self + Cross
Upsampling model	7B params	N/A	N/A	N/A	N/A

Table 1: Comparison of Model Specifications.

2.3 Text Conditioning

The choice of a text conditioning method plays a critical role in ensuring that the model accurately interprets and generates content based on textual input. To achieve robust text-to-image and text-to-video synthesis, we employ several strategies inspired by recent advancements in the field.

Utilizing Pre-Trained Text Encoders [GLIDE](#) [25] pioneered the learning of a transformer-based text encoder integrated with a denoising U-Net, setting a foundational approach for subsequent models. [Imagen](#) [26] and [ImagenVideo](#) [27] recommended using a pre-trained text encoder to condition diffusion models, a practice that has been widely adopted in subsequent works, including the use of a learnable projection layer on the input text embeddings to refine feature mapping.

Consistent with the approaches in [Imagen](#) [26], [DALL-E-3](#) [28], and [Pixart- \$\alpha\$](#) [8], we utilize the [T5-XXL](#) [29] text encoder to generate initial text embeddings. This choice is motivated by the success of these models in leveraging pre-trained text encoders to enhance the semantic understanding of the input text.

Cross-attention The two common architectures for conditioning diffusion transformers on text embedding tokens are *cross attention* ([Pixart- \$\alpha\$](#) [8], [MovieGen](#) [2]) and MM-DiT (introduced by [SD3](#) [12] and used by [FLUX.1](#) [30], [AuraFlow](#) [31], [CogVideoX](#) [3]). In MM-DiT, text embeddings

are processed in parallel to image patch embeddings. Unified attention layers facilitate the exchange of information between these two types of embeddings, suggesting that part of the transformation applied to image embeddings could be decoupled from the noise and timestep parameters. We use cross attention as we found it to work better than MM-DiT.

2.4 Image Conditioning

In many content-creation workflows, video generation starts from a given first frame, either real or generated, which we wish to animate and extend based on the text prompt. Different approaches exist for injecting the image-conditioning signal into the model ([SVD](#) [32], [LFDM](#) [33], [I2VGen-XL](#) [34], [Open-Sora Plan](#) [4]), typically requiring special tokens and models trained specifically for the image-to-video task. We base on and extend the approach presented in [Open-Sora](#) [35], which leverages the diffusion timestep as a conditioning indicator, allowing seamless conditioning on any part of the video.

In [DiT](#) [9] and [Pixart- \$\alpha\$](#) [8], the same timestep embedding is always injected to all tokens via AdaLN and all tokens are noised to the corresponding noise level. We relax this restriction by allowing a different timestep and corresponding noising level per token.

To train the model for first-frame conditioning, we occasionally set the timestep of the tokens belonging to the first video frame to a small random value and noise these tokens to the corresponding level. The model quickly learns to utilize this new information (when provided) as a conditioning signal.

During inference, the conditioning image is encoded into a latent tensor with a temporal dimension of 1 using our causal VAE encoder. This tensor is concatenated with random noise latents and flattened to form the initial set of tokens. The per-token denoising timesteps are set to a small value t_c for conditioning tokens and to $t = 1$ for all other tokens. Figure 9 illustrates this process.

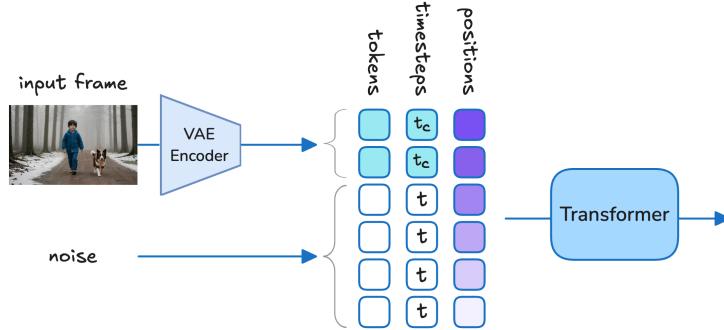


Figure 9: *LTX-Video* image-to-video inference pipeline – first-frame conditioning. The diffusion timestep and corresponding noise level is defined per-token. For example, conditioning tokens can have diffusion timesteps of $t_c = 0$ and contain un-noised encoded tokens.

2.5 Rectified-Flow Training

In this section, we describe specific design choices for the training procedure and the loss. These design choices significantly affect the training time and the resulting model quality.

2.5.1 Background

In *Rectified Flow* ([36], [SD3](#) [12]), the clean input z_0 is noised linearly in the forward process by the timestep $t \in [0, 1]$, according to $z_t = (1 - t)z_0 + t\epsilon$, where the noise ϵ is sampled from the standard-normal distribution $\mathcal{N}(0, I)$.

During training, t is randomly sampled from some distribution (uniform in the original diffusion paper – $t \sim \mathcal{U}(0, 1)$). In the original diffusion setting, the model was tasked with predicting the noise ϵ . Since this prediction task is not evenly-difficult, [SD3](#) [12] propose to task the network with predicting the *velocity* $v = \epsilon - z_0$ instead.

During inference, initial pure noise z_1 is gradually denoised towards a clean image z_0 . At each step, the denoising process is $z_{t-\Delta t} = z_t - \Delta t v_t^\theta$, where v_t^θ is the velocity predicted by the model at timestep t .

2.5.2 Timestep scheduling

SD3 [12] proposed sampling the diffusion timestep t during training from the *log-normal* distribution, replacing the uniform distribution $\mathcal{U}(t)$ used in earlier models. The motivation is to assign more training steps to timesteps in which the velocity-prediction task is more difficult. **SimpleDiffusion** [11] shows that at higher image resolutions, a higher level of noising is required to maintain the SNR. We adopt this recommendation and shift the timestep scheduler towards the higher-noise regions, depending on the number of tokens. To prevent starvation at the tail of the resolution we clamp the *pdf* at percentiles 0.5 and 99.9.

See Fig 10 for two timestep distribution shifting for two different resolutions.

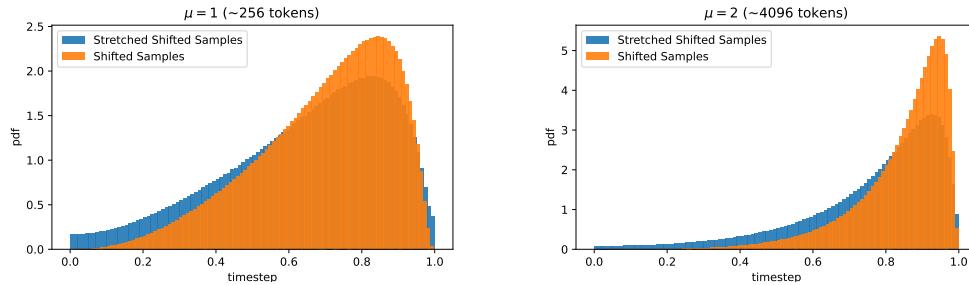


Figure 10: Timestep t sampling distribution during training, shown with two shifting parameters μ . We use the distributions shown in blue, which prevent near-zero probability at the tails.

2.5.3 Multi-resolution Training

To enable the model to generate videos at various resolutions and durations, we trained simultaneously on multiple combinations of resolutions and durations. We observed that after being exposed to a diverse range of width, height, and duration combinations, the model generalized well to unseen configurations.

During training, we ensured that all input samples contained approximately the same number of tokens by resizing the original videos to comparable token counts. To fix token counts across all sequences, we applied stochastic token dropping at rates ranging from 0% to 20%. This simple and efficient approach eliminated the need for complex token-packing or padding strategies while preserving diversity in the training data.

2.5.4 Training on Images

We include image training alongside regular video training, treating it as one of the resolution-duration combinations. Image datasets can enrich the set of concepts encountered during training, incorporating those not typically present in video datasets.

3 Data Preparation

Our training dataset comprises a robust collection of publicly available data, supplemented with licensed material, ensuring a diverse and comprehensive training environment and enabling our model to generate a wide variety of visual content. This section describes our data processing pipeline – see Fig 11.

Quality Control and Filtering Quality control is a critical aspect of our data preparation process. We train and employ an aesthetic model to evaluate both videos and images. The model was trained on tens of thousands of image pairs, manually tagged to identify the superior image in terms of aesthetic quality.

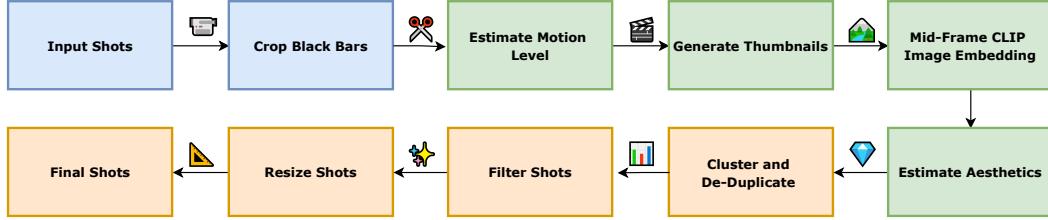


Figure 11: Our data processing pipeline.

To sample pairs for manual tagging, we labeled millions of samples using a multi-labeling network and sampled only pairs that share at least one of the top three labels. This approach helped minimize distribution shifts when filtering data based on aesthetics.

The training data was used to train a [Siamese Network \[37\]](#) designed to predict an aesthetic score that preserves the order relations established by the tagged pairs. Once trained, the model computes an aesthetic score for each sample, and we filter out those with scores below a certain threshold.

This process ensures that we select only the most visually appealing content, which is crucial for training our models to generate high-quality outputs.

Motion and Aspect Ratio Filtering In addition to aesthetic filtering, we actively remove videos that exhibit insignificant motion to ensure that our dataset focuses on dynamic content, which is more relevant to our model’s target capabilities. Videos are also processed to crop out any black bars, thus standardizing the aspect ratios and enhancing the usable visual data.

Fine-Tuning with Aesthetic Content For fine-tuning, we selectively use the most aesthetic content identified by our filtering processes. This approach helps in achieving more visually appealing results in the generated outputs, aligning with high standards of image and video quality.

See Fig 14b for the distribution of the clip durations in our filtered data.

Captioning and Metadata Enhancement To enhance the metadata of our training data, we utilize an internal automatic image and video captioner to re-caption the entire training set. This re-captioning process ensures that the text descriptions are accurate and relevant, providing better context for the training models and improving the alignment between visual content and textual annotations. Examples of input videos and generated captions are shown in Fig 12. Fig 13 shows a word-cloud visualizing the distribution of caption words in our entire data, and the distribution of the number of words per captions is shown in Fig 14a.



A large, colorful dragon puppet is carried down a street by a crowd of people. The dragon puppet is predominantly white with blue, green, and red accents, and it has large, expressive eyes and a long, serpentine body; it is suspended from a pole held by a person in the crowd, and its head moves slightly as it is moved down the street. The crowd is composed of people of all ages, many of whom are wearing winter clothes; they are cheering and waving their arms in the air, and some are taking pictures or videos with their phones. The street is lined with buildings, some of which have signs in Chinese characters; there are also a few trees and streetlights visible. The camera follows the dragon puppet from behind as it moves down the street, and it pans slightly to capture the surrounding crowd and the buildings; the lighting is natural and slightly overcast, casting soft shadows on the street. The scene is captured in real-life footage.



A Dalmatian dog stands on a window ledge, looks around, and then jumps. The dog has black spots on its white fur and is standing on a window ledge of a yellow building with green shutters. The dog looks to the left and then jumps from the window ledge. The camera is positioned across the street from the dog and remains stationary throughout the video. The lighting is natural sunlight, and the colors in the video are bright and vibrant. The scene is captured in real-life footage.

Figure 12: Our video captioner – input video (generated by [Sora \[1\]](#)) and our generated captions.



Figure 13: A word cloud of our captioned training data.

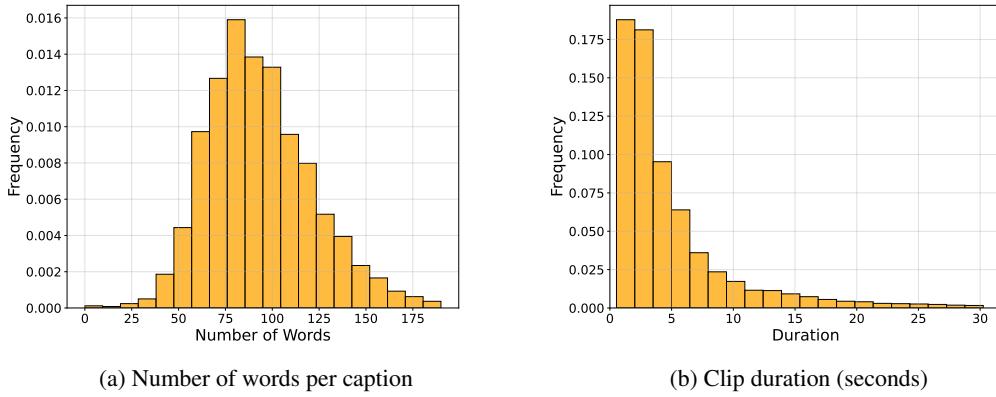


Figure 14: Training data statistics.

4 Experiments

4.1 Training

We trained our model using the *ADAM-W* optimizer. After pre-training, we fine-tuned our model on a subset of the data containing high-aesthetic videos.

4.2 Evaluation

Following MovieGen [2], we conducted a human survey to evaluate the quality of *LTX-Video* compared to current state-of-the-art models of similar size: Open-Sora Plan [4], CogVideoX [3] (2B) and PyramidFlow [5]. We used 1,000 prompts for the text-to-video task and 1,000 pairs of images and prompts for the image-to-video task (the images were generated using FLUX.1 [30]). We then generated 5-second videos at resolution of 768×512 using all evaluated models with their default configurations. All videos were generated using 40 diffusion steps.

The survey involved 20 participants, each of whom was shown each time a randomly ordered pair of videos from a randomly selected pair of evaluated models (not necessarily including *LTX-Video*). The model names were not revealed to the participants. For each pair, the videos were generated using the same prompt (and the same initial frame for the image-to-video task). Participants ranked the videos based on their overall preference, considering visual quality, motion fidelity, and prompt adherence. Participants could skip voting on a pair if they were unsure, using a provided 'skip' option. To ensure that our survey included a sufficient number of evaluators, we divided the participants into two groups of 10 and found that the win rates between groups differed by less than 2%.

The results for each category were summarized as the percentage of tests in which each model won (calculated as $wins/(wins + losses)$). The survey results are presented in Table 2. As shown, *LTX-Video* significantly outperforms other similar-sized models, even with its considerable speed advantage. Figure 15 illustrates the pairwise win ratios. Since we evaluated all model pairs (not just *LTX-Video* against others), we present the full pairwise results.

Table 2: Survey results showing the percentage of tests in which each model won for text-to-video and image-to-video tasks.

	Open-Sora Plan	CogVideoX 2B	PyramidFlow	LTX-Video (Ours)
Text-to-video	20%	38%	51%	85%
Image-to-video	20%	47%	35%	91%

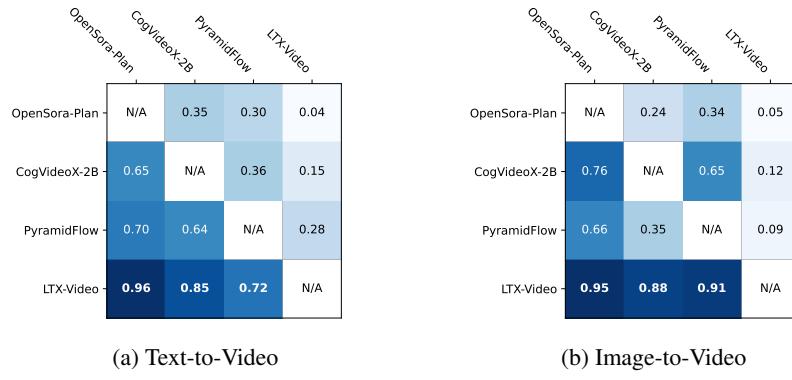


Figure 15: Pairwise performance matrix: each row indicates the win ratios of a model against all other models.

4.3 Ablations

4.3.1 Reconstruction GAN vs. traditional GAN

The high compression rate employed by our VAE challenges the reconstruction of high-frequency details, particularly in video frames that combine fast motion with intricate detail. Figure 16 illustrates such an example. At a high compression rate of 1:192, the commonly used combination of reconstruction and GAN losses fails to achieve consistent reconstruction of the original video (in this example, artifacts are visible in a single frame, but more commonly, inconsistencies become evident only when the video is played). Our proposed *Reconstruction GAN* loss significantly reduces these visible artifacts.



Figure 16: VAE reconstruction example – input frame (left), reconstruction with standard GAN loss (center), improved reconstruction with our proposed *Reconstruction GAN* loss (right)

4.3.2 RoPE frequency spacing

As discussed in Sec 2.2.1, we compared two variants for the RoPE frequency spacing – exponential and inverse-exponential. Fig 17 shows the training loss of two diffusion training experiments with

the two frequency spacing options. The experiments were conducted using the same architecture, hyper-parameters and data that was used in our *LTX-Video* training. As can be seen, the loss with inverse-exponential spacing remains higher.

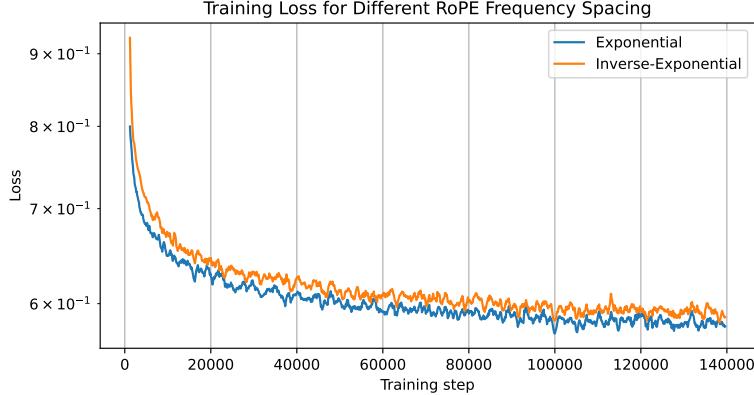


Figure 17: Video diffusion transformer training with two RoPE frequency spacing options - exponential and inverse-exponential. The loss is consistently better with exponential spacing.

4.4 Denoising VAE Decoder

Our holistic approach tasks the VAE decoder with performing the last denoising step in conjunction with converting the latents into pixels. To validate this design choice, we performed an internal user study comparing videos generated according to our approach to videos generated with the common approach, where denoising is performed solely by the diffusion-transformer, in latent space.

For the first set of results, our VAE decoder was conditioned on timestep $t = 0.05$. For the second set, the VAE decoder was conditioned on timestep $t = 0.0$ and did not perform any denoising.

The survey results indicated that videos generated by our method were strongly preferred over the standard results. The improvement was particularly evident in high-motion videos, where artifacts caused by the strong compression were mitigated by the VAE decoder’s last-step denoising.

5 Limitations

Model Sensitivity to Prompt Formulation While *LTX-Video* demonstrates strong prompt adherence, its performance can vary significantly based on the quality and clarity of textual prompts. Poorly formulated or very ambiguous prompts might result in less coherent outputs.

Limited Support for Long Videos Currently, the model focuses on generating short videos of up to ten seconds. Extending the architecture to support longer durations while maintaining temporal consistency and prompt fidelity remains an open area for future research.

Domain-Specific Generalization The model’s ability to adapt to domain-specific tasks (e.g., multi-view synthesis or fine-grained editing) has not been extensively tested, and further experimentation is required to assess its performance in specialized applications.

6 Social Impact

Accessibility and Democratization Our model is designed with accessibility in mind. Unlike many state-of-the-art text-to-video models that require high-end hardware and significant computational resources, our model is optimized for efficiency and can run on consumer-grade GPUs. This design choice makes advanced text-to-video generation more accessible to researchers, developers, and enthusiasts who may not have access to expensive hardware setups.

Open-Source Contribution By releasing our model as open-source, we aim to foster innovation and collaboration in the AI community. Open access to the model encourages diverse applications, including educational tools, creative content generation, and rapid prototyping for small and medium-sized enterprises, which may lack resources for training large models.

Environmental Considerations The relatively small size of our model not only reduces the hardware requirements but also lowers the energy consumption associated with training and deployment. This contributes to a more sustainable approach to deploying AI technologies.

Potential Risks and Mitigation While our model lowers barriers to entry, we recognize that making such technologies widely available could pose risks, such as misuse for generating misleading content. To mitigate this, we have included clear guidelines and disclaimers in our documentation to encourage responsible use.

7 Conclusion

In this paper, we introduced *LTX-Video*, a state-of-the-art transformer-based latent diffusion model designed for text-to-video and image-to-video generation. By addressing key limitations of existing methods, such as constrained temporal modeling and inefficient spatial compression, *LTX-Video* achieves faster-than-real-time generation while maintaining high motion fidelity, temporal consistency, and strong alignment with input prompts or conditioning frames.

At the core of *LTX-Video* is a holistic approach to latent diffusion, which seamlessly integrates the Video-VAE and the denoising transformer. This integration is achieved by relocating the patchifying operation from the transformer’s input to the VAE encoder, enabling efficient processing within a compressed latent space. Furthermore, the model introduces a novel shared diffusion objective between the VAE decoder and the transformer, effectively fusing the final diffusion step with the latent-to-pixel decoding stage. This innovation ensures fine-detail generation without the need for additional upsampling modules.

LTX-Video sets a new benchmark for text-to-video generation, outperforming SOTA open-source models of similar scale in speed and quality. Its ability to efficiently generate high-resolution videos while preserving coherence and adherence to prompts underscores the potential of latent diffusion models for video generation tasks.

The accessibility of *LTX-Video* further amplifies its impact, as its efficient design allows it to run on consumer-grade GPUs, democratizing advanced video generation capabilities. By lowering hardware requirements, *LTX-Video* opens the door to researchers, developers, and creative professionals who may not have access to high-end compute resources.

Future work may explore extending *LTX-Video*’s architecture to support longer videos, incorporate advanced temporal coherence techniques, and investigate its adaptability to domain-specific tasks such as multi-view video synthesis or fine-grained motion editing. By enabling faster, high-quality outputs with lower compute demands, *LTX-Video* represents a significant step forward in creative content creation, accessible AI, and scalable video modeling.

A young woman in a traditional Mongolian dress is peeking through a sheer white curtain, her face showing a mix of curiosity and apprehension. The woman has long black hair styled in two braids, adorned with white beads, and her eyes are wide with a hint of surprise. Her dress is a vibrant blue with intricate gold embroidery, and she wears a matching headband with a similar design...



A woman with long brown hair and bangs sits in the driver's seat of a car, smiling slightly as she looks to her right. She is wearing a black jacket over a white shirt. The camera is positioned outside the car, looking in through the driver's side window. The car is dark-colored and the background is blurred, suggesting a residential street...



A woman with blood on her face and a white tank top looks down and to her right, then back up as she speaks. She has dark hair pulled back, light skin, and her face and chest are covered in blood. The camera angle is a close-up, focused on the woman's face and upper torso. The lighting is dim and blue-toned, creating a somber and intense atmosphere. The scene appears to be from a movie or TV show.



A man in a dimly lit room talks on a vintage telephone, hangs up, and looks down with a sad expression. He holds the black rotary phone to his right ear with his right hand, his left hand holding a rocks glass with amber liquid. He wears a brown suit jacket over a white shirt, and a gold ring on his left ring finger. His short hair is neatly combed, and he has light skin with visible wrinkles around his eyes. The camera...



A man in a suit enters a room and speaks to two women sitting on a couch. The man, wearing a dark suit with a gold tie, enters the room from the left and walks towards the center of the frame. He has short gray hair, light skin, and a serious expression. He places his right hand on the back of a chair as he approaches the couch. Two women are seated on a light-colored couch in the background. The woman on the left wears a light blue sweater and has short blonde hair. The scene appears to be from a film or television show.



A young woman in a classroom receives a phone call, listens intently, and hangs up. The woman, with dark skin and long black hair styled in braids, wears a white turtleneck sweater, an orange cardigan, and a silver chain necklace; she holds a white smartphone to her right ear, her expression shifting from a smile to concern as she listens; she raises her eyebrows...



A man walks towards a window, looks out, and then turns around. He has short, dark hair, dark skin, and is wearing a brown coat over a red and gray scarf. He walks from left to right towards a window, his gaze fixed on something outside. The camera...



Figure 18: Additional text-to-video samples generated by *LTX-Video*, highlighting our model's high level of prompt adherence, visual quality and motion fidelity. Each row shows evenly-spaced frames from a generated 5-second video. Prompt adherence is indicated in green.

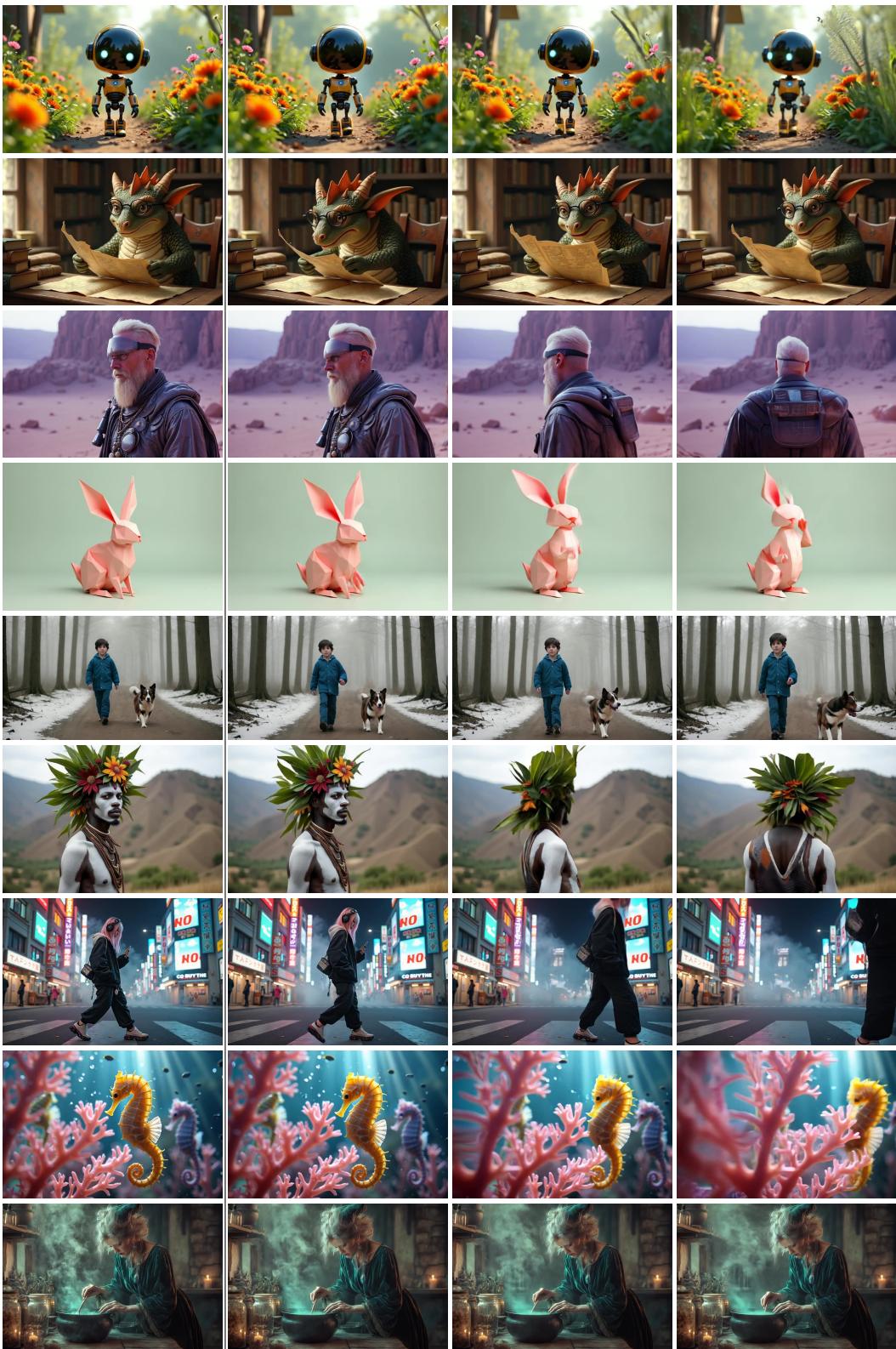


Figure 19: Additional image-to-video results. Video generation is conditioned on the first frame (left column) and on the text prompt (not shown).

References

- [1] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>, 3, 2024.
- [2] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.
- [3] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
- [4] PKU-Yuan Lab and Tuzhan AI. Open-sora-plan. *Github*, 2024. <https://doi.org/10.5281/zenodo.10948109>.
- [5] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.
- [6] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models, 2024.
- [7] Christina Zhang, Simran Motwani, Matthew Yu, Ji Hou, Felix Juefei-Xu, Sam Tsai, Peter Vajda, Zijian He, and Jialiang Wang. Pixel-space post-training of latent diffusion models. *arXiv preprint arXiv:2409.17565*, 2024.
- [8] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.
- [9] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.
- [10] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [11] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023.
- [12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.
- [13] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanyvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- [14] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.
- [16] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- [17] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

- [18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- [19] Zeyu Lu, Zidong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and Lei Bai. Fit: Flexible vision transformer for diffusion model. *arXiv preprint arXiv:2402.12376*, 2024.
- [20] Alpha VLLM. Large diffusion transformer. *GitHub*, 2024.
- [21] Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In *Forty-first International Conference on Machine Learning*, 2024.
- [22] Federico Barbero, Alex Vitvitskyi, Christos Perivolopoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*, 2024.
- [23] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.
- [24] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pages arXiv–1607, 2016.
- [25] 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.
- [26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [27] 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.
- [28] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.
- [29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [30] Black Forest Labs. Flux.1. *GitHub*, 2024. <https://github.com/black-forest-labs/flux>.
- [31] fofr. Auraflow v0.1, an open exploration of large rectified flow models. *GitHub*, 2024. <https://github.com/fofr/cog-aura-flow>.
- [32] 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.
- [33] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18444–18455, 2023.
- [34] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

- [35] HPC-AI Tech. Open-sora. *GitHub*, 2024. <https://github.com/hpcaitech/Open-Sora>.
- [36] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [37] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a ‘siamese’ time delay neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 6, pages 737–744, 1993.