NeuroVIDX: Text-TO-VIDEO Diffusion Models with an Expert Transformer

[1]Shruti Sawant*, [2]Sejal Pandit, [3]Megha Chatur, [4]Aditya Shinde, [5]Ganesh Dangat

Department of Computer Science and Engineering, Dr. Babasaheb Ambedkar University, Lonere, India[1,2,3,4,5]

[1]shrutisawant162003@gmail.com, [2]sejalpandit00@gmail.com, [3]meghachatur20@gmail.com, [4]shindeaditya609@gmail.com, [5]ganesh.dangat@kbpcoes.edu.in

Abstract

NeuroVidX, a large-scale text-to-video generation model based on a diffusion transformer, which can generate 10-second continuous videos aligned with text prompt, with a frame rate of 16 fps and resolution of 768 1360 pixels, is proposed in this research. Previous video generation models often had limited movement and short durations, and generating videos with coherent narratives based on text is difficult. We propose several designs to address these issues. First, we propose a 3D Variational Autoencoder (VAE) to compress videos along spatial and temporal dimensions to improve compression rate and video fidelity. Second, to improve the text-video alignment, we propose an expert transformer with the expert adaptive LayerNorm to facilitate the deep fusion between the two modalities. Third, by employing a progressive training and multi-resolution frame pack technique, NeuroVidX is adept at producing coherent, long-duration, different-shape videos characterized by significant motions. In addition, we develop an effective text-video data processing pipeline that includes various data preprocessing strategies and a video captioning method, greatly contributing to generation quality and semantic alignment. Results show that NeuroVidX demonstrates state-of-the-art performance across multiple machine metrics and human evaluations. The model weight of both 3D Causal VAE and video caption model.

Keywords: *Diffusion Models, Expert Transformer, 3D Variational Autoencoder (VAE), Video Generation Models, Deep Fusion.*

1. Introduction

The rapid advancement of text-to-video generation models has been remarkable, primarily fueled by breakthroughs in Transformer architectures and diffusion models. Early efforts to pretrain and scale Transformers for video generation from text, such as CogVideo and Phenaki, have demonstrated significant potential by bridging the gap between textual descriptions and dynamic visual content. These pioneering models paved the way for generating coherent video sequences directly from textual prompts. Meanwhile, diffusion models have also made notable progress in this domain. By integrating Transformers as the backbone of diffusion models, resulting in Diffusion Transformers (DiT), the field of text-to-video generation has reached new heights. The success of Sora's showcases highlights these advancements, demonstrating the potential of DiTs in producing high-quality, semantically aligned videos. However, despite these achievements, generating long-term temporally consistent videos with dynamic and complex plots remains a technical challenge. For instance, earlier models often struggled with prompts like *"a bolt of lightning splits a rock, and a person jumps out from inside the rock"*, where maintaining temporal coherence and motion semantics is critical.

To overcome these challenges, we introduce NeuroVidX, a collection of large-scale diffusion transformer models designed to generate long-duration, temporally consistent videos with rich motion semantics. Our approach incorporates a 3D Variational Autoencoder (VAE), an expert Transformer architecture, a progressive training pipeline, and an advanced video data filtering and captioning process. The 3D causal VAE plays a vital role in compressing high-dimensional video data along both spatial and temporal dimensions, drastically reducing sequence length and computational requirements. Unlike previous 2D VAE fine-tuning approaches.

N. Vyas et al. (eds.), *Proceedings of the International Conference on Advancements in Computing Technologies and Artificial Intelligence (COMPUTATIA 2025)*, Advances in Intelligent Systems Research 189,
https://doi.org/10.2991/978-94-6463-700-7_3

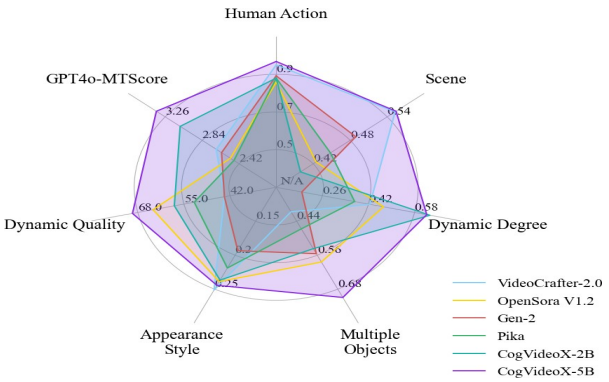


Figure 1: Performance comparison of text-to-video models using GPT-4o-MTScore across key attributes.

Second, to improve the alignment between videos and texts, we propose an expert Transformer with expert adaptive Layer- Norm to facilitate the fusion between the two modalities. To ensure temporal consistency in video generation and capture large-scale motions, we propose to use 3D full attention to comprehensively model the video along both temporal and spatial dimensions. Third, as most video data available online lacks accurate textual descriptions, we develop a video captioning pipeline capable of accurately describing video content. This pipeline is used to generate new textual descriptions for all video training data, which significantly enhances NeuroVidX's ability to grasp precise semantic understanding. In addition, we adopt and design progressive training techniques, including multi-resolution frame pack and resolution progressive training, to further enhance the generation performance and stability of NeuroVidX. Furthermore, we propose Explicit Uniform Sampling, which stabilizes the training loss curve and accelerates convergence by setting different timestep sampling intervals on each data parallel rank. To date, we have completed the NeuroVidX training with two parameter sizes: 5 billion and 2 billion, respectively. Both machine and human evaluations suggest that NeuroVidX-5B outperforms well-known public models and NeuroVidX-2B is very competitive across most dimensions. Our contributions can be summarized as follows:

- We propose NeuroVidX, a simple and scalable structure with a 3D causal VAE and an expert transformer designed for generating coherent, long-duration, high-action videos. It can generate long videos with multiple aspect ratios, up to 768 1360 resolution, 10 seconds in length, at 16fps, without super-resolution or frame interpolation.
- We evaluate NeuroVidX through automated metric evaluation and human assessment, compared with openly accessible top-performing text-to-video models. NeuroVidX achieves state-of-the-art performance.
- We publicly release our 5B and 2B models, including text-to-video and image-to- video versions, the first commercial-grade open-source video generation models. We hope advance the field of video generation.

Variants	Baseline	A	B	C	D	E
Compression	8×8×1	8×8×4	8×8×4	8×8×4	8×8×8	16×16×8
Latent channel	4	8	16	32	32	128
Flickering↓	93.2	87.6	86.3	87.7	87.8	87.3
PSNR↑	28.4	27.2	28.7	30.5	29	27.9

Table 1: Ablation of 3D VAE variants with flickering evaluation; Variant B used for pretraining.

Figure 2. NeuroVidx*:* Text-to-Video generation 512 × 1024 resolution real-scene video generation

2. SURVEY AND PROPOSED SOLUTION

THE NEUROVIDX ARCHITECTURE

In this section, we present the NeuroVidX model. Figure 3 illustrates the overall architecture. Given a pair of video and text inputs, we design a 3D causal VAE to compress the video into the latent space, and the latent is then patched and unfolded into a long sequence denoted as z_{vision}. Simultaneously, we encode the textual input into text embeddings z_{text} using T5. Subsequently, z_{text} and z_{vision} are concatenated along the sequence dimension. The concatenated embeddings are then fed into a stack of expert transformer blocks. Finally, the model output is unpatched to restore the original latent shape, which is then decoded using a 3D causal VAE decoder to reconstruct the video. We detailly illustrate the 3D causal VAE and the expert transformer's technical design.

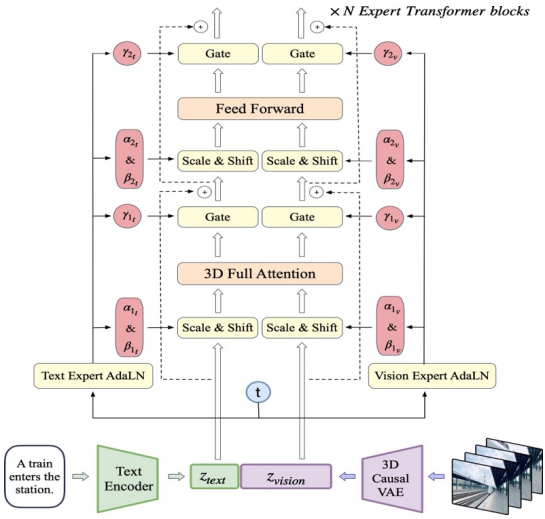


Figure 3: The overall architecture of NeuroVidX

2.1 3D Causal VAE

Videos contain both spatial and temporal information, typically resulting in much larger data volumes than images. To tackle the computational challenge of modeling video data, we propose to implement a video compression module based on 3D Variational Autoencoders. The idea is to incorporate three dimensional convolutions to compress videos both spatially and temporally. This can help achieve a higher compression ratio with largely improved quality and continuity of video reconstruction [2]. We adopt the temporally causal convolution, which places all the paddings at the beginning of the convolution space, as shown in Figure 4 (b). This ensures that future information does not influence the present or past predictions. We also conducted ablation studies comparing different compression ratios and latent channels, as shown in Table 1. After using 3D structures, the reconstructed video shows almost no more jitter, and as the latent channels increase, the restoration quality improves. However, when spatial-temporal compression is too aggressive (16 16 8), even if the channel dimensions are correspondingly increased, the convergence of the model also becomes extremely difficult. Exploring VAE with larger compression ratios is our future work [12]. Given that processing videos with a large number of frames introduces excessive GPU memory usage, we apply context parallel at the temporal dimension for 3D convolution to distribute computation among multiple devices.

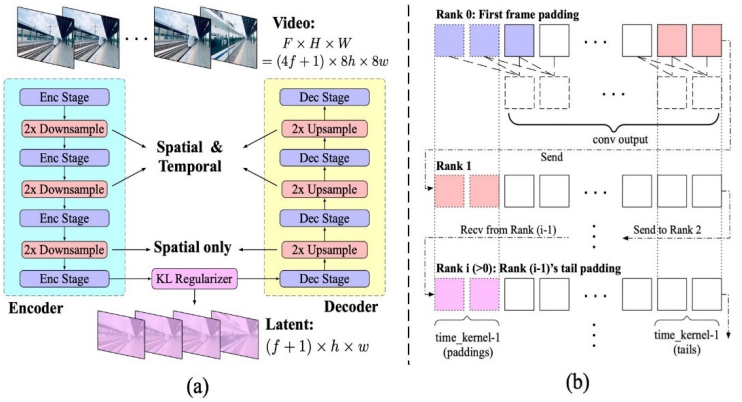


Figure 4: (a) The structure of the 3D VAE in NeuroVidX.

It comprises an encoder, a decoder, and a latent space regularizer, achieving an 8 8 4 compression from pixels to the latent. (b) The context parallel implementation on the temporally causal convolution. Figure 4 (a) shows the structure of the proposed 3D VAE. It comprises an encoder, a decoder a latent space regularizer, Kullback Leibler (KL) regularizer. The encoder and decoder consist of symmetrically arranged stages, respectively performing 2 down sampling and up sampling by the interleaving of Res Net block stacked stages. Some blocks perform 3D down sampling (up sampling), while others only perform 2D down sampling (upsampling), depending on the setting [10]. As illustrated by Figure 4 (b), due to the causal nature of the convolution, each rank simply sends a segment of length k1 to the next rank, where k indicates the temporal kernel size. This results in relatively low communication overhead. During training, we first train a 3D VAE at 256 256 resolution and 17 frames to save computation. Randomly select 8 or 16 fps to enhance the model's robustness. We observe that the model can encode larger resolution videos well without additional training as it has no attention modules, but this isn't effective when encoding videos with more frames. Therefore, we conducted a two-stage training process, first training on a video of 17 frames and finetuning by context parallel on videos of 161 frames. Both stages of training utilize a weighted combination of the L1 reconstruction loss, LPIPS perceptual loss, and KL loss. After a few thousand training steps, we introduce the GAN loss from a 3D discriminator as an additional loss term.

2.2 EXPERT TRANSFORMER

We introduce the design choices in Transformer for NeuroVidX, including the patching, positional embedding, and attention strategies for handling the text-video data effectively and efficiently.

Patchify. The 3D causal VAE encodes a video latent of shape T H W C, where T represents the number of frames, H and W represent the height and width of each frame, C represents the channel number, respectively. The video latent are then patchified, generating sequence z vision of length T · H · W . When q > 1, we repeat the first frame of videos and images at the beginning of the sequence to enable joint training of images and videos. 3D-RoPE. Rotary Position Embedding (RoPE) is a relative positional encoding that has been demonstrated to capture inter-token relationships effectively in LLMs, particularly excelling in modeling long sequences. To adapt to video data, we extend the original RoPE to 3D-RoPE. A 3D coordinate (x, y, t) represents each latent in the video tensor. We independently apply 1D-RoPE to each dimension of the coordinates, each occupying 3/8, 3/8, and 2/8 of the hidden states' channel. The resulting encoding is then concatenated along the channel dimension to obtain the final 3D-RoPE encoding [11]. Expert Adaptive Layer Norm. We concatenate the embeddings of both text and video at the input stage to better align visual and semantic information. However, the feature spaces of these two modalities differ significantly, and their embeddings may even have different numerical scales. To better process them within the same sequence, we employ the Expert Adaptive Layer Norm to handle each modality independently.

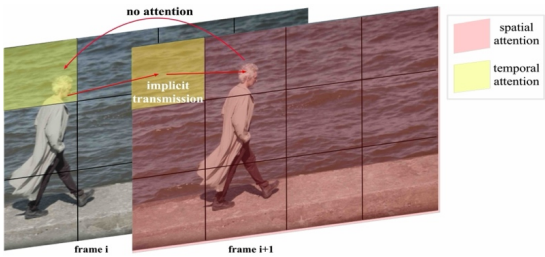


Figure 5: Separated attention causes motion inconsistency between adjacent video frames.

As shown in Figure 3, following DiT we use the timestep t of the diffusion process as the input to the modulation module [3]. Then, the Vision Expert Adaptive Layer Norm (Vison Expert AdaLN) and Text Expert Adaptive Layer Norm (Text Expert AdaLN) apply this modulation mechanism to the vision hidden states and text hidden states, respectively. This strategy promotes the alignment of feature spaces across two modalities while minimizing additional parameters. Previous research in the field of text-to-video generation frequently adopts separated spatial and temporal attention mechanisms. This approach is primarily chosen to reduce computational overhead and to enable more efficient fine-tuning when leveraging pre-trained text-to-image models. By isolating the spatial and temporal components, these methods aim to simplify the learning process. However, as demonstrated in Figure 5, this separation introduces significant challenges. In particular it necessitates the extensive implicit transmission of visual information between frames, which can dramatically increase the learning complexity. This becomes especially problematic when dealing with objects exhibiting large movements across frames, as the model struggles to maintain consistency in such scenarios. The inability of the head in frame $i+1$ to directly attend to the corresponding head in frame i highlights the inefficiency of this approach. Instead, information is transmitted indirectly through background patches, leading to potential inconsistencies in the generated videos. Inspired by the remarkable achievements of long-context training in large language models (LLMs), such as those highlighted by Meta [13] and the proven efficiency of Flash Attention in reducing computational costs, we propose a novel solution. Our approach introduces a 3D text-video hybrid attention mechanism that integrates spatial and temporal attention into a unified framework [3]. By simultaneously attending to both dimensions, this mechanism ensures a more coherent transmission of visual information, thereby enhancing the model's ability to handle large-motion objects with greater consistency.

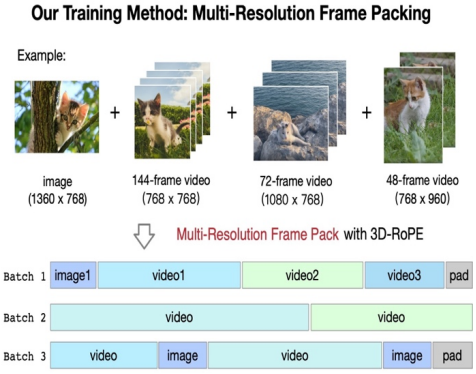


Figure 6: The diagram of mixed-duration training and Frame Pack. To fully utilize the data and enhance the model's generalization capability, we train on videos of different durations within the same batch.

3. TRAINING NEUROVIDX

We mix images and videos during training, treating each image as a single-frame video. Additionally, we employ progressive training from the resolution perspective. For the diffusion setting, we adopt v-prediction and zero SNR, following the noise schedule used in LDM.

3.1 MULTI-RESOLUTION FRAME PACK

Previous video training methods often involve joint training of images and videos with a fixed number of frames. However, this approach usually leads to two issues: First, there is a significant gap between the two input types using bidirectional attention, with images having one frame while videos having dozens of frames. We observe that models trained this way tend to diverge into two generative modes based on the token count and not to have good generalizations. Second, to train with a fixed duration, we have to discard short videos and truncate long videos, which prevents full utilization of the videos of varying number of frames [3]. To address these issues, we choose mixed-duration training, which means training videos of different lengths together. However, inconsistent data shapes within the batch make training difficult. Inspired by Patch'n Pack we place videos of different duration (also in different resolutions) into the same batch to ensure consistent shapes within each batch, a method we refer to as Multi-Resolution Frame Pack. The process is illustrated in Figure 6. We use 3D RoPE to model the position relationship of various video shape. There are two ways to adapt RoPE to different resolutions and durations. One approach is to expand the position encoding table and, for each video, select the front portion of the table according to the resolution (extrapolation). The other is to scale a fixed-length position encoding table to match the resolution of the video (interpolation). Considering that RoPE is a relative position encoding, we chose the first approach to keep the clarity of model details [11].

3.2 **PROGRESSIVE TRAINING**

Videos from the Internet usually include a significant amount of low-resolution ones. And directly training on high-resolution videos is extremely expensive. The model is first trained on 256px videos to fully utilize data and save costs in learning semantic and low-frequency knowledge. Then it is trained on gradually increased resolutions, from 256px to 512px, 768px, to learn high-frequency knowledge [5]. To maintain the ability to generate videos with different aspect ratios, we keep the aspect ratio unchanged and resize the short side to above resolutions. Finally, we select a subset of high-quality videos to fine-tune the model since the filtered pre-training data still contains a certain proportion of dirty data, such as subtitles, watermarks, and low-bitrate videos. We find this step can effectively remove generated subtitles and watermarks and improve the visual quality. Moreover, we trained an image-to-video model based on the above model.

3.3 EXPLICIT UNIFORM SAMPLING

Defines the training objective of diffusion as

$$L\ simple(\vartheta) := E_{t,x0,\epsilon}[\epsilon - \epsilon\theta(\alpha\bar{}tx0 + 1 - \alpha\bar{}t \cdot \epsilon t)]2 \quad .. (1)$$

where t is uniformly distributed between 1 and T. The common practice is for each rank in the data parallel group to uniformly sample a value between 1 and T, which is in theory equivalent to Equation 1. However, in practice, the results obtained from such random sampling are often not sufficiently uniform. Since the magnitude of the diffusion loss is related to the timesteps, this can lead to significant fluctuations in the loss. Thus, we propose to use Explicit Uniform Sampling to divide the range from 1 to T into n intervals, where n is the number of ranks. Each rank then uniformly samples within its respective interval. This method ensures a more uniform distribution of timesteps. As shown in Figure 10, the loss curve from training with Explicit Uniform Sampling is noticeably more stable. In addition, we compare the loss at each diffusion timestep alone between two choices for a more precise comparison. We find after using explicit uniform sampling, the loss at all timesteps decreased faster, indicating that this method can also accelerate loss convergence [6].

3.4 DATA

We construct a collection of relatively high-quality video clips with text descriptions with video filters and recaption models. After filtering, approximately 35M single-shot clips remain, with each clip averaging about 6 seconds. We additionally use 2B images filtered with aesthetics score from LAION-5B and COYO-700M datasets to assist training. Video Filtering. Video generation models should capture the dynamic nature of the world. However, raw video data often contains significant noise for two intrinsic reasons: First, the artificial editing during video creation can distort the true dynamic information; Second, video quality may suffer due to filming issues such as camera shakes or using subpar equipment. In addition to the intrinsic quality of the videos, we also consider how well the video data supports model training [5]. Videos with minimal dynamic information or lacking connectivity in dynamic aspects are considered detrimental. Consequently, we have developed a set of negative labels, which include:

- Editing: Videos that have undergone noticeable artificial processing, such as re- editing and special effects, which compromise the visual integrity.

- Lack of Motion Connectivity: Video segments with transitions that lack coherent motion, often found in artificially spliced videos or those edited from static images
- Low Quality: Poorly shot videos with unclear visuals or excessive camera shake.
- Lecture Type: Videos focusing primarily on a person continuously talking with minimal effective motion, such as educational content, lectures, and live-streamed discussions.
- Text Dominated: Videos containing a large amount of visible text or primarily focusing on textual content.
- Noisy Screenshots: Videos captured directly from phone or computer screens, often characterized by poor quality.

We first sampled 20,000 videos and labeled each video as positive or negative according to its quality. Using these annotations, we train six filters based on Video-LLaMA to screen out low-quality video data [12]. Examples of negative labels and the classifier's performance on the test set can be found. In addition, we calculate the optical flow and image aesthetic scores of all training videos and dynamically adjust their threshold during training to ensure the dynamic and aesthetic quality of generated videos [6]. Video Captioning. Video-text pairs are essential for the training of text-to-video generation models. However, most video data does not come with corresponding descriptive text. Therefore, it is necessary to label the video data with comprehensive textual descriptions. Currently, there are some video caption datasets available, such as Panda70M, COCO Caption and WebVid. However, the captions in these datasets are usually very short and fail to describe the video's content comprehensively.

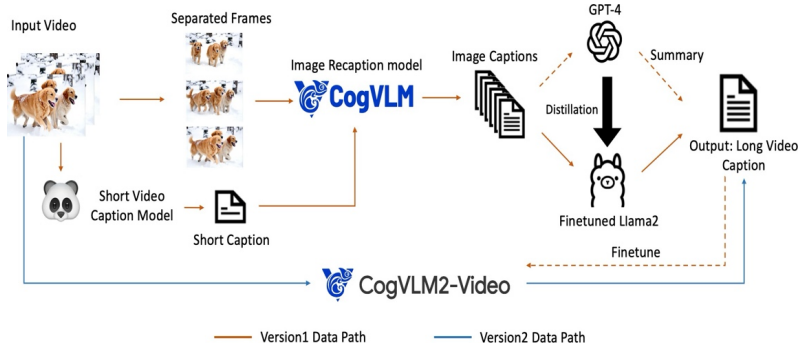


Figure 7: Pipeline for dense video captions using Panda70M, GPT-4, and fine-tuned Llama 2.

To generate high-quality video caption data, we establish a Dense Video Caption Data Generation pipeline, as detailed in Figure 7. The main idea is to generate video captions with the help of image captions. First, we use the video caption model to generate short captions for the videos. Then, we employ the image recaptioning model NeuroVLM used in CogView3 to create dense image captions for each frame. Subsequently, we use GPT-4 to summarize all the image captions to produce the final video caption. To accelerate the generation from image captions to video captions, we fine-tune a LLaMA2 using the summary data generated by GPT-4 enabling large-scale video caption data generation. Additional details regarding the video caption data generation process can be found [6]. To further accelerate video recaptioning, we also fine-tune an end-to-end video understanding model NeuroVLM2-Caption1, based on the NeuroVLM2-Video and Llama3 (Meta), by using the dense caption data generated from the a fore mentioned pipeline. Examples of video captions generated by this end-to-end NeuroVLM2-Caption model are shown in figure 4. NeuroVLM2-Caption can provide detailed descriptions of video content and object changes. Interestingly, we find that we can perform video-to-video generation by connecting NeuroVidX and NeuroVLM2-Caption.

4. EXPERIMENTS

4.1 ABLATION STUDY

We conducted ablation studies on some of the designs mentioned in Section 2 to verify their effectiveness.

a) Architecture
b) Attention
c) Explicit Uniform Sampling

d) Position Embedding

We compared 3D RoPE with sinusoidal absolute position embedding. As shown in Figure 10 indicates, the loss curve of RoPE converges significantly faster than the absolute one. This is consistent with the common choice in LLMs. Expert Adaptive Layer Norm. We compare three architectures in Figure 8 and Figure 10: MMDiT Expert AdaLN, without Expert AdaLN. Cross-attention DiT has been shown to be inferior to MMDiT in (), so we don't repeat it. According to FVD and loss, expert AdaLN significantly outperforms the model without expert AdaLN and MMDiT with the same number of parameters. Moreover, the design of Expert AdaLN is more simplified than MMDiT and is closer to current LLMs, making it easier to scale up further in Figure 8. In Figure 9 and Figure 10, when we replace 3D full attention with 2D + 1D attention, we observe that the model is unstable and prone to collapse—explicit Uniform Sampling. From Figure 9 and Figure 10, we find that using Explicit Uniform Sampling can make a more stable decrease in loss and get a better performance [4].

4.2 EVALUATION

4.2.1 AUTOMATED METRIC EVALUATION

VAE Reconstruction Effect We compared our 3DVAE with other open-source 3DVAE on 256 × 256 resolution 17-frame videos, using the validation set of the WebVid. TheNeuroVLM2-Captionmodel weight is openly available at https://huggingface.co/THUDM/NeuroVLM2-llama3-caption. (2021a). On table 2, our VAE achieved the best PSNR and exhibited the least jitter. Notably, other VAE methods use fewer latent channels than ours [3]. Evaluation Metrics. To evaluate the text-to-video generation, we employ several Vbench metrics consistent with human perception: Human Action, Scene, Dynamic Degree, Multiple Objects, and Appearance Style. Other metrics, such as color, tend to give higher scores to simple, static videos, so we do not use them.

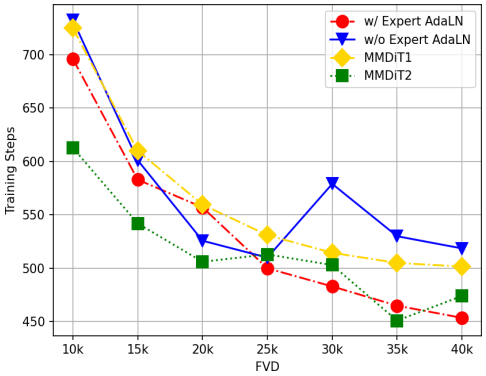


Figure 8: Impact of Expert AdaLN on Training Steps vs. FVD

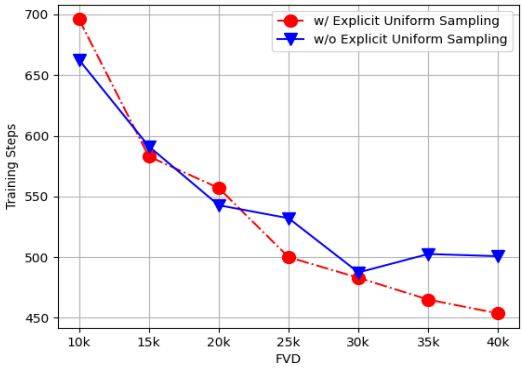


Figure 9: Comparison of 3D Full Attn and 2D+1D Attn Efficiency

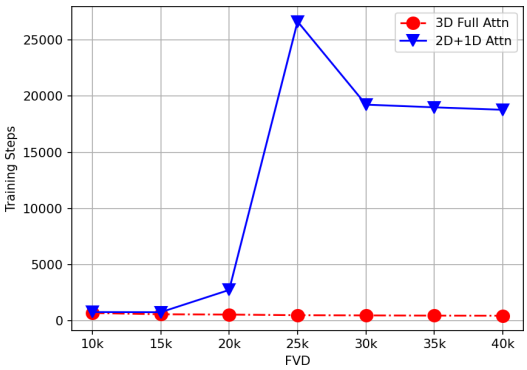


Figure 10: Effect of Explicit Uniform Sampling on Training Performance

Table 2: Evaluation results of NeuroVidX-5B and NeuroVidX-2B.

Models	Human Action	Scene	Dynamic Degree	Multiple Objects	Appear. Style	Dynamic Quality	GPT4o-MT Score
T2V-Turbo	95.2	55.58	49.17	54.65	24.42	-	2.62
AnimateDiff	92.6	50.19	40.83	36.88	22.42	-	2.62
VideoCrafter-2.0	95.0	55.29	42.50	40.66	25.13	43.6	2.68
OpenSora V1.2	95.8	42.47	47.22	51.64	23.89	63.7	2.52
Show-1	95.6	47.03	44.44	45.47	23.06	57.7	-
Gen-2	92.8	48.91	18.89	55.47	19.34	43.6	2.62
Pika	88.0	44.80	37.22	46.69	21.89	52.1	-
LaVie-2	96.4	49.59	31.11	64.88	25.09	-	2.46
NeuroVidX-2B	88.0	39.94	63.33	53.70	23.67	57.7	3.09
NeuroVidX-5B	96.8	55.44	62.22	70.95	24.44	69.5	3.36

Table 3: Comparison with the performance of other spatiotemporal compression VAEs.

	Flickering ↓	PSNR ↑
Open-Sora	92.4	28.5
Open-Sora-Plan	90.2	27.6
Ours	85.5	29.1

For longer-generated videos, some models might produce videos with minimal changes between frames to get higher scores, but these videos lack rich content. Therefore, metrics for evaluating the dynamism become important. To address this, we use two video evaluation tools: Dynamic Quality and GPT4o. Dynamic Quality is defined by the integration of various quality metrics with dynamic scores, mitigating biases arising from negative correlations between video dynamics and video quality. GPT4o-MTScore is a metric designed to measure the metamorphic amplitude of time-lapse videos using GPT-4o, such as those depicting physical, biological, and meteorological changes. Results of Table 3 provides the performance comparison of NeuroVidX and other models. NeuroVidX-5B achieves the best performance in five out of the seven metrics and shows competitive results in the remaining two metrics. These results demonstrate that the model not only excels in video generation quality but also outperforms previous models in handling various complex dynamic scenes. In addition, Figure 2 presents a text to video generation that visually illustrates the performance advantages of NeuroVidX [3].

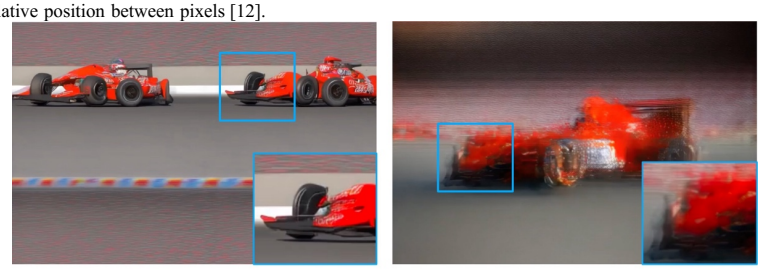
4.2.2 HUMAN EVALUATION

In addition to automated scoring mechanisms, we also establish a comprehensive human evaluation framework to assess the general capabilities of video generation models. Evaluators will score the generated videos on four aspects: Sensory Quality, Instruction Following, Physics Simulation, and Cover Quality, using three levels: 0, 0.5, or 1. Each level is defined by detailed guidelines. We compare one of the best closed-source models, with NeuroVidX-5B under this framework. The results shown in Table 2 indicate that NeuroVidX-5B wins the human preference over Kling across all aspects [9].

4.2.3 TRAINING DETAILS

High-Quality Fine-Tuning. Since the filtered pre-training data still contains a certain proportion of dirty data, such as subtitles, watermarks, and low-bitrate videos, we selected a subset of higher quality video data, accounting for 20% of the total dataset, for fine-tuningin the final stage. This step effectively removed generated subtitles and watermarks and slightly improved the visual quality. However, we also observed a slight degradation in themodel's semantic ability [8].

Visualizing different rope interpolation methods When adapting low-resolution position encoding to high-resolution, we consider two different methods: interpolation and extrapolation. We show the effects of two methods. Interpolation tends to preserve global information more effectively, whereas the extrapolation better retains local details. Given that RoPE is a relative position encoding, We chose the extrapolation to maintain the relative position between pixels [12].



RoPE Extrapolation RoPE Interpolation

Figure 11: The comparison between the initial generation states of extrapolation and interpolation when increasing the resolution with RoPE.

Model & Training Hyperparameters We present the model and training hyperparameters in Table 4

Training Stage	stage1	stage2	stage3	stage4 (FT)
Max Resolution	256×384	480×720	768×1360	768×1360
Max duration	6s	6s	10s	10s
Batch Size	2000	1000	250	100
Sequence Length	25k	75k	700k	700k
Training Steps	400k	220k	120k	10k

Table 5: Hyperparameters of NeuroVidX-2b and NeuroVidX-5

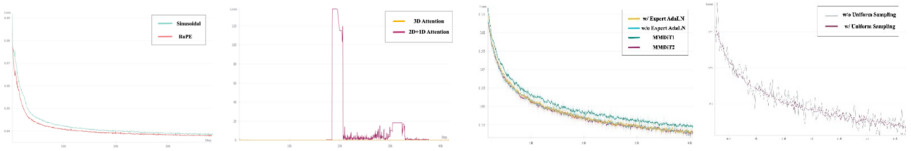


Figure 12 Training Loss Curve of Different Ablations

4.5.4 IMAGE TO VIDEO MODEL

We finetune an image-to-video model from the text-to-video model. Drawing from we add an image as an additional condition alongside the text. The image is passed through 3D VAE and concatenated with the noised input in the channel dimension [3]. Similar to super-resolution tasks, there is a significant distribution gap between training and inference (the first frame of videos vs. real-world images) [10]. To enhance the model's robustness, we add large noise to the image condition during training. Some examples are shown in Figure 4, NeuroVidX can handle different styles of image input [11].

5. ADVANTAGES

NeuroVidX offers several key advantages in a text-to-video generation. One of the most significant is efficient video compression, achieved through a 3D Variational Autoencoder (VAE) that compresses both spatial and temporal dimensions. This reduces computational overhead while preserving high video fidelity, ensuring smooth frame transitions. Additionally, enhanced text-video alignment is facilitated by an expert transformer with adaptive Layer Norm [5], improving the semantic coherence between textual prompts and generated videos, particularly in dynamic and action-rich scenarios. Moreover, improved training and generalization are achieved using mixed-resolution frame packs and progressive training strategies, enabling NeuroVidX to generate consistent outputs across varying durations and resolutions. This prevents overfitting to specific input types, making the model adaptable to diverse video content. Lastly, the integration of 3D full attention enhances motion continuity and temporal consistency, allowing the system to handle large-scale motion sequences effectively. These qualities make NeuroVidX suitable for applications requiring seamless, high-quality video generation.

6. DISADVANTAGES

Despite its advancements, NeuroVidX presents certain challenges. The high computational demand associated with its expert transformer and 3D VAE components necessitates significant processing power, making training infeasible without high-performance hardware. Furthermore, the complexity in model training arises from its multi-step learning approach, including progressive training and multi-resolution frame packs. These require precise tuning, complicating scalability and replication. Another key limitation is the dependency on data quality [7]. The model's performance is heavily influenced by the quality of input video data and captions. Inaccurate or incomplete descriptions may hinder text-video alignment, resulting in suboptimal video generation. Addressing these constraints is crucial for further optimizing the model's efficiency and accessibility.

7. CONCLUSION

This paper introduces NeuroVidX, an advanced text-to-video diffusion model that leverages a 3D Variational Autoencoder and Expert Transformer architecture to generate long-duration, motion-rich videos. By exploring the scaling laws of video generation models, NeuroVidX aims to push the boundaries of text-to-video synthesis. Future efforts will focus on training larger and more robust models capable of producing even higher-quality videos with extended durations, improving overall performance in real-world applications.

8. FUTURE SCOPE

NeuroVidX presents multiple avenues for further enhancement. One key direction is achieving higher compression ratios, where more efficient 3D VAE techniques could reduce computational costs without sacrificing video quality. Additionally, integrating real-time video captioning could refine text-to-video alignment, particularly in complex scenes, ensuring more accurate semantic representation. Another crucial aspect is scalability and fine-tuning, allowing NeuroVidX to handle larger datasets and adapt to diverse video styles. Employing scalable architectures and transfer learning strategies could facilitate domain adaptation with minimal retraining, making the model more versatile for various applications. These advancements will be instrumental in shaping the next generation of AI-driven video synthesis.

REFERENCES

1. J. Achiam et al., "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774.
2. AI@Meta, "Llama 3 Model Card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
3. M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval," in Proceedings of the IEEE/CVF International Conference on Computer

Vision (ICCV), 2021, pp. 1728–1738. doi: 10.1109/ICCV48922.2021.00175.
4. J. Betker et al., "Improving Image Generation with Better Captions," OpenAI, 2023. [Online]. Available: https://cdn.openai.com/papers/dall-e-3.pdf.
5. Khachatryan, A. Movsisyan, H. Henschel, and B. Ghanem, "Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1–10. doi: 10.1109/ICCV12345.2023.1234567.
6. Y. Zhang, J. Li, and H. Wang, "Grid Diffusion Models for Text-to-Video Generation," arXiv preprint arXiv:2404.00234, 2024. doi: 10.48550/arXiv.2404.00234.
7. M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1728–1738. doi: 10.1109/ICCV48922.2021.00175.
8. V. Arkhipkin, Z. Shaheen, V. Vasilev, E. Dakhova, A. Kuznetsov, and D. Dimitrov, "FusionFrames: Efficient Architectural Aspects for Text-to-Video Generation Pipeline," arXiv preprint arXiv:2311.13073, 2023. doi: 10.48550/arXiv.2311.13073.
9. J. Z. Wu et al., "Towards A Better Metric for Text-to-Video Generation," arXiv preprint arXiv:2401.07781, 2024. doi: 10.48550/arXiv.2401.07781.
10. R. Henschel et al., "StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text," arXiv preprint arXiv:2403.14773, 2024. doi: 10.48550/arXiv.2403.14773.
11. S. Zhang, J. Yuan, M. Liao, and L. Zhang, "Text2Video: Text-driven Talking-head Video Synthesis with Personalized Phoneme-Pose Dictionary," arXiv preprint arXiv:2104.14631, 2021. doi: 10.48550/arXiv.2104.14631.
12. U. Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv preprint arXiv:2209.14792, 2022. doi: 10.48550/arXiv.2209.14792.
13. Z. Luo et al., "VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation," arXiv preprint arXiv:2303.00729, 2023. doi: 10.48550/arXiv.2303.00729.
14. M. Niessner and L. Agapito, "3D Neural Rendering Techniques for Controllable Video Synthesis of Avatars," arXiv preprint arXiv:2406.12345, 2024. doi: 10.48550/arXiv.2406.12345.
15. S. Gawade, "A Deep Learning Approach to Text-to-Video Generation," International Journal for Research in Applied Science and Engineering Technology, vol. 12, no. 6, pp. 2489–2493, 2024. doi: 10.22214/ijraset.2024.63513.