One-Minute Video Generation with Test-Time Training

Karan Dalal*[4] Daniel Koceja*[2] Jiarui Xu*[1,3] Yue Zhao[5] Shihao Han[1]
Ka Chun Cheung[1] Jan Kautz[1] Yejin Choi[1] Yu Sun[1,2] Xiaolong Wang[1,3]
[1]NVIDIA [2]Stanford University [3]UCSD [4]UC Berkeley [5]UT Austin

Figure 1. TTT layers enable pre-trained Diffusion Transformers to generate one-minute videos from text storyboards. We use *Tom and Jerry* cartoons as a proof-of-concept. The videos tell complete stories with coherent scenes composed of dynamic motion. Every video is produced directly by the model in a single shot, without editing, stitching, or post-processing. All stories are original.

Abstract

Transformers today still struggle to generate one-minute videos because self-attention layers are inefficient for long context. Alternatives such as Mamba layers struggle to produce coherent scenes because their hidden states are small and less expressive. We experiment with Test-Time Training (TTT) layers, whose hidden states themselves can be neural networks, therefore larger and more expressive. Adding TTT layers into a pre-trained Transformer enables it to generate one-minute videos from text storyboards. We curate a dataset based on Tom and Jerry cartoons as a proof-of-concept benchmark. Compared to baselines such as Mamba 2, Gated DeltaNet, and sliding-window attention layers, TTT layers generate much more coherent videos that tell complete stories, leading by 34 Elo points in a human evaluation of 100 videos per method. Although promising, our results are still limited in physical realism, and the efficiency of our implementation can be further improved.

Sample videos, code and annotations are available at: https://test-time-training.github.io/video-dit

1. Introduction

Despite the remarkable progress in visual and physical realism, state-of-the-art video Transformers are still generating mostly short clips of single scenes. At the time of writing (March 2025), the maximum length of public video generation APIs is 8 seconds for Veo 2 (Google), 20 seconds for Sora (OpenAI), 16 seconds for MovieGen (Meta), and 10 seconds for Ray 2 (Luma).

A major challenge in video generation is long context, because the computation and memory cost of self-attention layers in Transformers increases quadratically with context length. This challenge is especially acute for videos with dynamic motion, whose context cannot be easily compressed by a tokenizer. Using a standard tokenizer, our one-minute videos requires over 300k tokens in context.

To address this challenge, recent work in long video

*Joint first authors

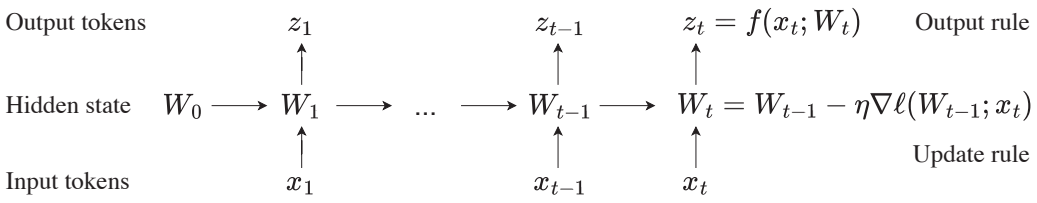

Figure 2. All RNN layers can be expressed as a hidden state that transitions according to an update rule. The key idea in [43] is to make the hidden state itself a model f with weights W, and the update rule a gradient step on the self-supervised loss ℓ. Therefore, updating the hidden state on a test sequence is equivalent to training the model f at test time. This process, known as test-time training (TTT), is programmed into TTT layers. Figure and caption taken from [43].

generation has explored (bi-directional) RNN layers as an efficient alternative to self-attention because their cost increases linearly with context length [47]. Modern RNN layers, especially variants of linear attention [23, 37] such as Mamba [8, 12] and DeltaNet [35, 52, 53], have shown impressive results for natural language tasks. However, we have yet to see RNN layers generating videos with highly dynamic motion spanning multiple scenes, let alone complex narratives.

We believe that these RNN layers end up generating less coherent videos because their hidden states are small and less expressive. RNN layers can only store past tokens into a hidden state of fixed size, which is only a matrix for linear attention variants such as Mamba and DeltaNet. It is inherently challenging to compress hundreds of thousands of vectors into a matrix with only thousands in rank. As a consequence, these RNN layers have a hard time remembering the relationships between distant tokens.

We experiment with an alternative class of RNN layers whose hidden states themselves can be neural networks. Specifically, we use two-layer MLPs with $2\times$ more hidden cells and richer nonlinearities than the matrix hidden states in linear attention variants. Since the neural network hidden states are updated by training even on test sequences, these new layers are called Test-Time Training (TTT) layers [43].

We start from a pre-trained Diffusion Transformer (CogVideo-X 5B [19]) that could only generate 3-second clips at 16 FPS (or 6-seconds at 8 FPS). Then, we simply add TTT layers initialized from scratch and fine-tune this modified model to generate one-minute videos from text storyboards. During fine-tuning, we limit the self-attention layers to 3-second segments so their cost stays manageable. With only preliminary systems optimization, our training run takes the equivalent of 50 hours on 256 H100s.

We curate a text-to-video dataset based on ≈ 7 hours of *Tom and Jerry* cartoons with human-annotated descriptions. We intentionally limit our scope to the specific domain of *Tom and Jerry* for fast research iteration. As a proof-of-concept, this dataset emphasizes complex, multi-scene, and long-range narratives with dynamic motion, where progress is still needed. It has less emphasis on visual and physical realism, where remarkable progress has already been made. We believe that improvements in long-context capabilities for this specific domain will transfer well to general-purpose video generation.

Compared to baselines such as Mamba 2, Gated DeltaNet, and sliding-window attention layers, TTT layers generate much more coherent videos that tell complete stories, leading by 34 Elo points in a human evaluation of 100 videos per method. For context, GPT-4 scores 46 Elo points over GPT-3.5 Turbo, and GPT-4o scores 29 over GPT-4 Turbo in Chatbot Arena [6].

Sample videos, code and annotations are available at: https://test-time-training.github.io/video-dit

2. Test-Time Training Layers

Following standard practice [44, 54], each video is pre-processed into a sequence of T tokens, where T is determined by its duration and resolution. This section reviews Test-Time Training (TTT) layers for general sequence modeling, using some of the exposition in Section 2 of [43]. We first discuss how to process general input sequences in a causal manner (chronological order). Section 3 discusses how to use RNN layers in a non-causal backbone by invoking them in opposite directions.

2.1. TTT as Updating a Hidden State

All RNN layers compress historical context in a hidden state of fixed size. This compression has two consequences. On one hand, mapping an input token x_t to output token z_t is efficient, because both the update rule and output rule take constant time per token. On the other hand, an RNN layer's ability to remember long context is limited by the amount of information its hidden state can store. The goal of [43] is to design RNN layers with expressive hidden states that can compress massive context. As an inspiration, they observe that self-supervised learning can compress a massive training set into the weights of a machine learning model.

The key idea in [43] is to use self-supervised learning to compress the historical context $x_1, \ldots, x_t$ into a hidden

state W_t, by making the context an unlabeled dataset and the hidden state the weights of a machine learning model f. The update rule, illustrated in Figure 2, is a step of gradient descent on some self-supervised loss ℓ:

$$W_t = W_{t-1} - \eta \nabla \ell(W_{t-1}; x_t), \tag{1}$$

with learning rate η. Intuitively, the output token is just the prediction on x_t, made by f with the updated weights W_t:

$$z_t = f(x_t; W_t). \tag{2}$$

One choice of ℓ is reconstructing x_t itself. To make the learning problem nontrivial, we first process x_t into a corrupted input $\tilde{x}_t$ (details in Subsection 2.2), then optimize:

$$\ell(W; x_t) = \|f(\tilde{x}_t; W) - x_t\|^2. \tag{3}$$

Similar to denoising autoencoders [46], f needs to discover the correlations between dimensions of x_t in order to reconstruct it from partial information $\tilde{x}_t$.

As with other RNN layers and self-attention, this algorithm that maps an input sequence $x_1, \ldots, x_T$ to output sequence $z_1, \ldots, z_T$ can be programmed into the forward pass of a sequence modeling layer. Even at test time, the layer still trains a different sequence of weights $W_1, \ldots, W_T$ for every input sequence. Therefore, it is called *Test-Time Training (TTT) layer*.

Conceptually, calling backward on $\nabla \ell$ means taking gradients of gradients – a well-explored technique in meta-learning. TTT layers have the same interface as RNN layers and self-attention, therefore can be replaced in any larger network architecture. [43] refers to training the larger network as the *outer loop*, and training W within each TTT layer as the *inner loop*.

2.2. Learning a Self-Supervised Task for TTT

Arguably, the most important part of TTT is the self-supervised task, because it determines the kind of features that W will learn from the input sequence. Instead of hand-crafting a self-supervised task from human priors, [43] takes a more end-to-end approach.

Concretely, [43] learns the self-supervised task as part of the outer loop. Starting from the naive reconstruction task in Equation 3, they add some outer-loop parameters to make this task learnable. In Subsection 2.1, we did not specify the corruption that produces $\tilde{x}_t$ from x_t. [43] uses low-rank projection $\tilde{x}_t = \theta_K x_t$, where θ_K is a learnable matrix.

Moreover, perhaps not all the information in x_t is worth remembering, so the reconstruction label can also be a low-rank projection $\theta_V x_t$ instead of x_t. In summary, the self-supervised loss in [43] is:

$$\ell(W; x_t) = \|f(\theta_K x_t; W) - \theta_V x_t\|^2. \tag{4}$$

Lastly, since $\theta_K x_t$ has fewer dimensions than x_t, [43] can no longer use the output rule in Equation 2. So they make another projection $\theta_Q x_t$, and change the output rule to:

$$z_t = f(\theta_Q x_t; W_t). \tag{5}$$

Note that in the inner loop, only W is optimized, therefore written as an argument of ℓ; the θs are "hyper-parameters" of this inner-loop loss function. $\theta_K, \theta_V, \theta_Q$ are optimized in the outer loop, analogous to the Query, Key, and Value parameters of self-attention.

2.3. TTT-MLP Instantiation

Following [43], we instantiate the inner-loop model f as a wrapper around $f_{\texttt{MLP}}$, a two-layer MLP similar to those in Transformers. Specifically, the hidden dimension is $4\times$ the input dimension, followed by a GELU activation [16]. For better stability during TTT, f always contains a Layer Norm and residual connection. That is,

$$f(x) = x + \texttt{LN}(f_{\texttt{MLP}}(x)).$$

A TTT layer with this f is called TTT-MLP, which is the default instantiation throughout this paper. In Section 4 we also instantiate TTT-Linear (the f above wrapping around a linear model) as a baseline.

3. Approach

At a high level, our approach simply adds TTT layers to a pre-trained Diffusion Transformer and fine-tunes it on long videos with text annotations. At a practical level, making this approach work involves many design choices.

3.1. Architecture

Pre-trained Diffusion Transformer. Our approach of adding TTT layers then fine-tuning can, in principle, work with any backbone architecture. We choose Diffusion Transformers [32] for our initial demonstration because it is the most popular architecture for video generation. Since the cost of pre-training a Diffusion Transformer on videos is prohibitive, we start from a pre-trained checkpoint called CogVideo-X 5B [19].

Gating. Given an input sequence $X = (x_1, \ldots, x_T)$ where each token $x_t \in \mathbb{R}^d$, a TTT layer produces an output sequence $Z = (z_1, \ldots, z_T) = \texttt{TTT}(X)$. Each $z_t \in \mathbb{R}^d$ follows the recurrence described by Equations 1, 4 and 5 in Section 2. Naively inserting TTT layers into a pre-trained network would dramatically worsen its predictions at the beginning of fine-tuning, when the TTT layers are randomly initialized. To avoid this degradation, we gate $\texttt{TTT}$ with a learned vector $\alpha \in \mathbb{R}^d$ following standard practice [1]:

$$\texttt{gate}(\texttt{TTT}, X; \alpha) = \tanh(\alpha) \otimes \texttt{TTT}(X) + X, \tag{6}$$

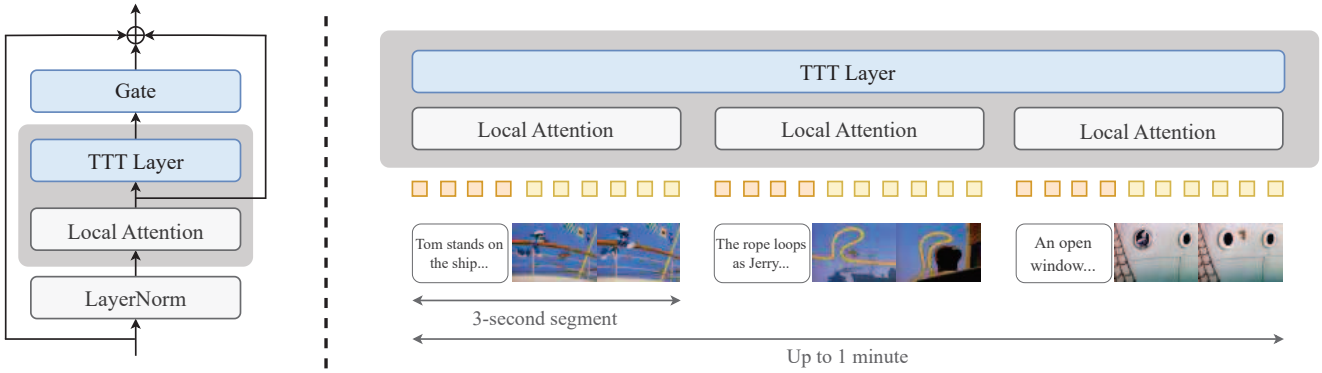

Figure 3. Overview of our approach. **Left**: Our modified architecture adds a TTT layer with a learnable gate after each attention layer. See Subsection 3.1. **Right**: Our overall pipeline creates input sequences composed of 3-second segments. This structure enables us to apply self-attention layers locally over segments and TTT layers globally over the entire sequence. See Subsection 3.2.

where $\tanh(\alpha) \in (-1, 1)^d$ is multiplied element-wise with each z_t in $Z = \texttt{TTT}(X)$. We initialize all values in α to 0.1, so the values in $\tanh(\alpha)$ are close to 0 (≈ 0.1) at the beginning of fine-tuning. This initialization of α allows TTT to still contribute to $\texttt{gate}(\texttt{TTT}, X; \alpha)$ without significantly overwriting X.

Bi-direction. Diffusion models, including CogVideo-X, are non-causal, meaning that an output token z_t can condition on all of $x_1, \ldots, x_T$ instead of only the past tokens $x_1, \ldots, x_t$. To use TTT layers in a non-causal manner, we apply a standard trick called bi-direction [30]. Given an operator $\texttt{rev}(X) = (x_T, \ldots, x_1)$ that reverses $X = (x_1, \ldots, x_T)$ in time, we define

$$\texttt{TTT}'(X) = \texttt{rev}(\texttt{TTT}(\texttt{rev}(X))). \tag{7}$$

Since rev is applied twice, $\texttt{TTT}'(X)$ is still in chronological order. But the TTT layer inside it now scans through X in reverse-chronological order.

Modified architecture. Standard Transformers, including CogVideo-X, contain interleaving sequence modeling blocks and MLP blocks. Specifically, a standard sequence modeling block takes an input sequence X and produces

$$X' = \texttt{self\_attn}(\texttt{LN}(X)) \tag{8}$$
$$Y = X' + X, \tag{9}$$

where LN is Layer Norm[1] and $X' + X$ forms a residual connection. We only modify the sequence modeling blocks, leaving everything else in the architecture unchanged. Each modified block, illustrated in the left panel of Figure 3, continues from the X' in Equation 8 and produces

$$Z = \texttt{gate}(\texttt{TTT}, X'; \alpha), \tag{10}$$
$$Z' = \texttt{gate}(\texttt{TTT}', Z; \beta), \tag{11}$$
$$Y = Z' + X. \tag{12}$$

[1]Diffusion Transformers such as CogVideo-X use adaptive LN [32].

Note that $\texttt{TTT}'$ only makes another call to TTT, so they share the same underlying parameters $\theta_K, \theta_V, \theta_Q$. But for gating, Equation 10 and 11 use different parameters α and β.

3.2. Overall Pipeline

In this subsection, we discuss how to create the input sequence of tokens to our architecture and how each sequence is processed in segments. Except for the first two text formats in the upcoming discussion, everything applies to both fine-tuning and inference. Our pipeline is illustrated in the right panel of Figure 3.

Scenes and segments. We structure our videos to contain multiple scenes,[2] and each scene contains one or more 3-second *segments*. We use a 3-second segment as the atomic unit of text-to-video pairing for three reasons:

- The maximum length of generation for the original pre-trained CogVideo-X is 3 seconds.
- The length of most scenes in the *Tom and Jerry* episodes is at least 3 seconds.
- Building a dataset with multiple stages (Subsection 3.3) is most convenient given 3-second segments.

Formats of text prompts. At inference time, a user can write the text prompt for a long video in any of the three formats listed below in increasing detail.

- **Format 1**: A short summary of the plot in 5-8 sentences. See examples in Figure 1.
- **Format 2**: A more detailed plot in roughly 20 sentences, with each sentence roughly corresponding to a 3-second segment. Sentences can be labeled as belonging to certain scenes or groups of scenes, but these labels will be treated only as suggestions.
- **Format 3**: A storyboard. Each 3-second segment is described by a paragraph of 3-5 sentences, containing

[2]A scene is loosely defined as "a part of a film in which the action happens in one place or is of one particular type." (Oxford Dictionary)

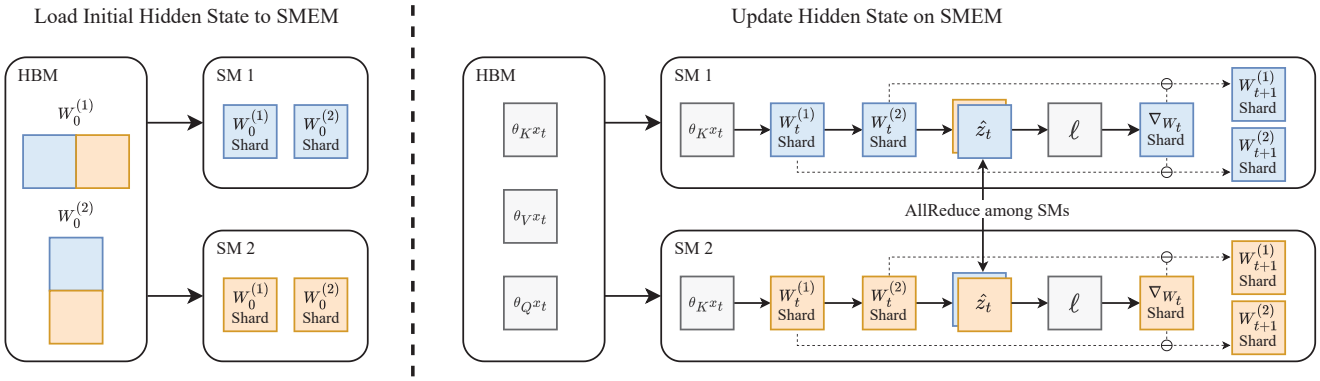

Figure 4. On-chip Tensor Parallel, discussed in Subsection 3.5. **Left:** To reduce the memory required on each SM for TTT-MLP, we shard the hidden state $W^{(1)}$ and $W^{(2)}$ across SMs, transferring them between HBM and SMEM only during initial loading and final output. **Right:** We update the hidden state entirely on-chip and use the DSMEM feature on the NVIDIA Hopper GPU architecture to `AllReduce` intermediate activations among SMs.

details such as background colors and camera movements. Groups of one or more paragraphs are strictly enforced as belonging to certain scenes with the keywords `<scene start>` and `<scene end>`.

See Appendix 6 for examples of each format. The actual input to our text tokenizer is always in Format 3 during both fine-tuning and inference. Conversion between the formats is performed by the GPT 4.5 API in the order of $1 \rightarrow 2 \rightarrow 3$.[3] For fine-tuning, our human annotations are already in Format 3, as discussed in Subsection 3.3.

From text to sequences. After the original CogVideo-X tokenizes the input text for each video, it concatenates the text tokens with 17,550 noisy video tokens. To generate a long video, we apply the same procedure independently for each 3-second segment. Specifically, given a storyboard in Format 3 with n paragraphs, we first produce n *sequence segments*, each containing text tokens extracted from the corresponding paragraph followed by 17,550 video tokens. Then we concatenate all m sequence segments together to form the input sequence, which now has interleaved text and video tokens.

Local attention, global TTT. CogVideo-X uses self-attention layers to process the entire input sequence globally for each video of maximum length 3 seconds, but global attention becomes inefficient for long videos. To avoid increasing the context length of self-attention layers, we make them local to each 3-second segment, attending to each of the n sequence segments independently.[4] The TTT layers process the entire input sequence globally because they are efficient in long context.

[3] We observe that converting from Format 1 directly to Format 3 results in worse ability to follow the style of the human annotations in Format 3 in the fine-tuning dataset.

[4] As an artifact of our pre-processing step, the sequence segments actually have an overlap of 1 latent frame (1350 tokens).

3.3. Fine-Tuning Recipe and Dataset

Multi-stage context extension. Following standard practice for LLMs [51], we extend the context length of our modified architecture to one minute in five stages. First, we fine-tune the entire pre-trained model on 3-second segments of *Tom and Jerry* to adapt it to this domain. New parameters (specifically those in TTT layers and gates) are assigned a higher learning rate during this stage. Over the next four stages, we fine-tune on videos of 9, 18, 30, and eventually 63 seconds. To avoid forgetting too much of the world knowledge from pre-training, we only fine-tune the TTT layers, gates, and self-attention layers, using a lower learning rate during these four stages. See Appendix A for the detailed recipe.

Super-resolution on original videos. We start with 81 episodes of *Tom and Jerry* released between 1940 and 1948. Each episode is about 5 minutes, adding up to about 7 hours for all episodes. The original videos vary in resolution, which is uniformly poor by modern standards. We run a video super-resolution model [49] on the original videos, producing visually enhanced videos with shared resolution of 720×480 for our dataset.

Multi-stage dataset. Following the structure discussed in Subsection 3.2, we first have human annotators break down each episode into scenes, then extract 3-second segments from each scene. Then, human annotators write a detailed paragraph for each 3-second segment.[5] Stage 1 fine-tunes directly on those segments. To create data for the last four stages, we concatenate contiguous 3-second segments into videos of 6, 18, 30 and 63 seconds together with their text

[5] Each paragraph includes 1–2 sentences describing the background, 1–2 sentences describing the characters, and 2 sentences describing actions and camera movements. On average, each paragraph contains 98 words, which corresponds to 132 tokens.

	Text Alignment	Motion Smoothness	Aesthetics	Scene Consistency	Average
Mamba 2	985	976	963	988	978
Gated DeltaNet	983	984	993	1004	991
Sliding Window	**1016**	1000	1006	975	999
TTT-MLP	1014	**1039**	**1037**	**1042**	**1033**

Table 1. Human evaluation results for one-minute video generation. TTT-MLP outperforms baselines by an average of +34 ELO points, notably in scene consistency (+38), motion smoothness (+39), and aesthetics (+31), demonstrating the advantage of its larger, expressive hidden state. For context, LMSys reports GPT-4 at +46 Elo points over GPT-3.5 Turbo, and GPT-4o at +29 over GPT-4 Turbo.

annotations. Scene boundaries are marked by the same keywords in Subsection 3.2. As a result, text annotations for all training videos are in Format 3.

3.4. Parallelization for Non-Causal Sequences

The update rule discussed in Section 2 cannot be naively parallelized across tokens in a sequence, since computing W_t requires $\nabla \ell(W_{t-1}; x_t)$, which in turn requires W_{t-1}. To enable parallelization, we update W on b tokens at a time, which [43] calls an inner-loop mini-batch. Throughout this paper, we set $b = 64$.

Concretely, for mini-batch $i = 1, \ldots, T/b$ (assuming T is an integer multiple of b),

$$W_{ib} = W_{(i-1)b} - \frac{\eta}{b} \sum_{t=(i-1)b+1}^{ib} \nabla \ell \left(W_{(i-1)b}; x_t \right). \quad (13)$$

Because the sequence is non-causal, we then use W_{ib} to produce the output tokens for all timesteps in mini-batch i:

$$z_t = f(W_{ib}; x_t), \qquad \text{for } t = (i-1)b + 1, \ldots, ib. \quad (14)$$

Note that $W_{(i-1)b+1}, \ldots, W_{ib-1}$ are no longer needed.

After this modification, f can process an (inner-loop) mini-batch of tokens in parallel, similar to how a regular MLP processes an (outer-loop) mini-batch of training data. As a side benefit, we observe that averaging gradients across tokens reduces variance and stabilizes each update to W.

3.5. On-Chip Tensor Parallel

Implementing TTT-MLP efficiently for GPUs requires special designs to take advantage of their memory hierarchy. A chip on a GPU is called a Streaming Multiprocessor (SM), analogous to a core on a CPU. All SMs on a GPU share a relatively slow but large global memory called HBM, then each SM has a fast but small on-chip memory called SMEM. Frequent data transfers between the SMEMs and HBM on a GPU can significantly hurt overall efficiency.

Both self-attention and Mamba have efficient implementations, such as Flash Attention [9], which use kernel fusion to minimize this kind of transfer. The high-level idea of these implementations is to load inputs and initial states into each SMEM, perform computations entirely on-chip, and write only the final outputs back to HBM. However, the hidden state for TTT-MLP, namely the weights $W^{(1)}$ and $W^{(2)}$ of the two-layer MLP f, is too large to be stored in the SMEM of a single SM (when combined with inputs and intermediate activations).

To reduce the memory required on each SM, we use Tensor Parallelism [39] to shard $W^{(1)}$ and $W^{(2)}$ across SMs, as shown in Figure 4. Similar to how large MLP layers can be sharded and trained across the HBMs of multiple GPUs, we apply the same idea now across the SMEMs of multiple SMs, treating each SM as the analogy of a GPU. We use the DSMEM feature on the NVIDIA Hopper GPU architecture to implement `AllReduce` among SMs. More details of our kernel are discussed in Appendix B.

Our implementation significantly improves efficiency, since hidden states and activations are now read from and written to HBMs only during initial loading and final output. As a general principle, if a model architecture f can be sharded with standard Tensor Parallelism across GPUs, then the same sharding strategy can be applied across SMs when f is used as the hidden state.

4. Experiments

We perform human evaluation on a multi-axis benchmark for TTT-MLP and five baselines with linear complexity: local attention, TTT-Linear, Mamba 2, Gated DeltaNet, and sliding window attention layers.

4.1. Baselines

Except for local attention, all baselines are added to the pre-trained Diffusion Transformer using the approach in Section 3.1; their modified architectures have 7.2B parameters. All baselines use the same training recipe, described in Appendix 2. Next we discuss the baselines in detail.

- **Local attention**: No modification to the original architecture, which performs local self-attention on each 3-second segment (17,550 video tokens) independently.
- **TTT-Linear** [43]: A TTT layer that instantiates $f(x) = x + \mathsf{LN}(f_{\mathsf{Linear}}(x))$, where f_{Linear} is a linear model.
- **Mamba 2** [8]: A modern RNN layer with a matrix hidden

Figure 5. Qualitative examples comparing TTT-MLP against Gated DeltaNet and Sliding-window attention. TTT-MLP demonstrates scene consistency by preserving details across transitions, and improved motion smoothness by depicting complex character actions.

state, which is $\approx 4\times$ larger than the hidden state in TTT-Linear but $\approx 2\times$ smaller than that in TTT-MLP.

- **Gated DeltaNet** [53]: An extension of DeltaNet [52] and Mamba 2 that uses the same hidden state with an improved update rule.
- **Sliding-window attention** [3]: Self-attention with a fixed window of 8192 tokens (about 1.5 seconds of video). Since CogVideo-X has 42 attention layers, their receptive field covers the entire one-minute context length.

4.2. Evaluation Axes and Protocol

For our multi-axes human evaluation, we adopt four axes from MovieGen [44]:[6]

- **Text following**: how well the generated video aligns with the given prompt.
- **Motion smoothness**: the viewing continuity between frames and the absence of visual artifacts.
- **Aesthetics**: visual quality of the generated videos, including lighting, colors, and camera effects.
- **Scene consistency**: consistency of characters and settings across scene transitions, including significant time gaps.

Evaluation protocol. We conduct evaluations for 18-second and one-minute videos, generating 100 videos per method for each length. All text prompts are produced by Claude 3.7 Sonnet. For the one-minute evaluation, we exclude the TTT-Linear and Local Attention baselines.

Since directly ranking multiple long videos would be challenging for evaluators, we instead rely on a pairwise comparison using an Elo-style rating system. Specifically, evaluators are presented with two videos and a criterion. They simply select the video that better meets the criterion.

To calculate Elo scores, we adopt the approach from LMSys ChatBot Arena [6], using the Bradley–Terry model with logistic regression to analyze pairwise comparisons.

4.3. Results

In this subsection, we discuss the results for one-minute evaluations, listed in Table 1. For discussion on 18-second evaluations, see Appendix 3.

On average, TTT-MLP outperforms Mamba 2, Gated DeltaNet, and Sliding Window Attention layers by +34 ELO points. Notably, TTT-MLP achieves +38 points on scene consistency, defined as the ability to preserve details across significant time gaps. We attribute this improvement to TTT-MLP's larger, expressive hidden state, enabling it to maintain context and connections across extended context lengths. To contextualize this gap, we refer to the LMSys Leaderboard:

Model	Elo
GPT-4o-2024-05-13	1285
GPT-4-Turbo-2024-04-09	1256
GPT-4-0613	1163
GPT-3.5-Turbo-0613	1117

GPT-4 is +46 Elo points above GPT-3.5 Turbo, while GPT-4o is +29 Elo points above GPT-4 Turbo, indicating that our differences are practically meaningful.

TTT-MLP also surpasses baselines by +39 points in motion smoothness and +31 points in aesthetics. This improvement largely stems from baselines degenerating in the latter half of generated videos, as illustrated by qualitative examples in Figure 5. For example, Sliding-window attention

[6]We omit "Realness" which does not apply to cartoons. We merge "Motion Naturalness" and "Motion Completeness" into "Motion Smoothness" to reduce evaluation cost. We adapt "Frame Consistency" to "Scene Consistency" for our multi-scene videos.

often exhibits abrupt and unnatural character movements towards the end of videos.

Finally, we find no significant differences among methods in prompt-following ability. This is expected since all methods utilize the same dense 3-second segment prompts and local attention over segment intervals (Section 3.2).

5. Related Work

Modern RNN layers, especially linear attention variants [23, 37], such as Mamba [8, 12] and DeltaNet [35, 52], have demonstrated impressive performance in natural language tasks. Inspired by their success and ideas from Fast Weight Programmers [7, 21, 24, 36], [43] proposes scalable and practical ways to make the hidden states large and nonlinear, therefore more expressive. Recent work [2] develops even larger and more nonlinear hidden states, and updates them with more sophisticated optimization techniques. The related work section in [43] contains a detailed discussion of inspirations for TTT layers. [48] gives a good overview of recent developments in RNN layers.

Long video modeling. Some early work [40] generates long videos by training GAN [11, 22] to predict the next frame based on the current frame and the motion vector. Generation quality has improved significantly due to recent progress in auto-regression (AR) and diffusion-based approaches [13, 25, 44, 54]. TATS [10] proposes the sliding window attention on the Transformer to generate videos longer than the training length. Phenaki [45] works in a similar auto-regressive way, but each frame is generated by MaskGIT [4]. Pre-trained diffusion models can be extended to generate longer videos by using cascade [15, 50, 55], streaming [17], and adding transitions [5].

Story synthesis methods such as [20, 26, 28, 29, 31, 33] generate sequences of images or videos corresponding to individual sentences in a text story. For example, Craft [14] generates videos of complex scenes through retrieval, and StoryDiffusion [56] uses diffusion to improve the smoothness of transitions between frames. While related to text-to-video generation, story synthesis methods usually need additional components in their pipeline to maintain coherence across scenes, which are not processed end-to-end.

6. Future Work

In this paper, we show how TTT layers enable the generation of coherent one-minute videos from text storyboards. We present a practical method for adding TTT layers into a pre-trained Diffusion Transformer, extending its generation capabilities from 3-three seconds up to one-a full minute. Our experimental results support the hypothesis that larger, more expressive hidden states better capture relationships between tokens and effectively model long context sequences. We outline several promising directions for future research:

Faster implementation. Our kernel for TTT-MLP can potentially leverage asynchronous operations to better hide latency and minimize register spills.

Better integration. Using bi-direction and learned gates is only one possible strategy for integrating TTT layers into a pre-trained model. Better strategies should further improve performance. Other video generation backbones, such as autoregressive models, might require different strategies.

Longer videos with larger hiddens states. Our approach can potentially be extended to generate much longer videos with linear complexity. The key to achieving that goal, we believe, is to instantiate the hidden states as much larger neural networks than our two-layer MLP. For example, f itself can be a Transformer.

Acknowledgements. We thank Hyperbolic Labs for compute support, Yuntian Deng for help with running experiments, and Aaryan Singhal, Arjun Vikram, and Ben Spector for help with systems questions. Yue Zhao would like to thank Philipp Krähenbühl for discussion and feedback. Yu Sun would like to thank his PhD advisor Alyosha Efros for the insightful advice of looking at the pixels when working on machine learning.

Note on authorship. Gashon Hussein and Youjin Song joined the team after an initial version of this project was submitted to CVPR, and have made major contributions to the final version. Due to CVPR policies preventing authorship changes after submission, their names could not appear on OpenReview or the conference website. However, we all agree that the official author list should include their names, as presented in our released PDFs. This project would not be possible without their work.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022. 3

[2] Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024. 8

[3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. 7

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. 8

[5] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion

model for generative transition and prediction. In *ICLR*, 2023. 8

[6] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. 2, 7

[7] Kevin Clark, Kelvin Guu, Ming-Wei Chang, Panupong Pasupat, Geoffrey Hinton, and Mohammad Norouzi. Meta-learning fast weight language models. *arXiv preprint arXiv:2212.02475*, 2022. 8

[8] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *ICML*, 2024. 2, 6, 8

[9] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *NeurIPS*, 2022. 6

[10] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, 2022. 8

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 8

[12] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM*, 2024. 2, 8

[13] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *ECCV*, 2024. 8

[14] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *ECCV*, 2018. 8

[15] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 8

[16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[17] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 8

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[19] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023. 2, 3

[20] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *NAACL*, 2016. 8

[21] Kazuki Irie, Imanol Schlag, Róbert Csordás, and Jürgen Schmidhuber. Going beyond linear transformers with recurrent fast weight programmers. *Advances in Neural Information Processing Systems*, 34:7703–7717, 2021. 8

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 8

[23] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, 2020. 2, 8

[24] Louis Kirsch and Jürgen Schmidhuber. Meta learning backpropagation and improving it. *NeurIPS*, 34:14122–14134, 2021. 8

[25] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv 2412.03603*, 2025. 8

[26] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *CVPR*, 2019. 8

[27] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *WACV*, 2024. 1

[28] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *CVPR*, 2024. 8

[29] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *ECCV*, 2022. 8

[30] Shentong Mo and Yapeng Tian. Scaling diffusion mamba with bidirectional ssms for efficient image and video generation, 2024. 4

[31] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *WACV*, 2024. 8

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *CVPR*, 2023. 3, 4

[33] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *CVPR*, 2023. 8

[34] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 1

[35] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *ICML*, 2021. 2, 8

[36] Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992. 8

[37] Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992. 2, 8

[38] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision, 2024. 1

[39] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019. 6

[40] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022. 8

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[42] Benjamin F Spector, Simran Arora, Aaryan Singhal, Daniel Y Fu, and Christopher Ré. Thunderkittens: Simple, fast, and adorable ai kernels. In *ICLR*, 2025. 1

[43] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024. 2, 3, 6, 8

[44] The Movie Gen team. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2, 7, 8

[45] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. In *ICLR*, 2023. 8

[46] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, page 1096–1103, 2008. 3

[47] Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu, Yaqiao Luo, Peizhao Zhang, Tingbo Hou, Peter Vajda, Niraj K. Jha, and Xiaoliang Dai. Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity, 2024. 2

[48] Ke Alexander Wang, Jiaxin Shi, and Emily B Fox. Test-time regression: a unifying framework for designing sequence models with associative memory. *arXiv preprint arXiv:2501.12352*, 2025. 8

[49] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, 2021. 5

[50] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *IJCV*, 2024. 8

[51] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*, 2023. 5

[52] Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *NeurIPS*, 2024. 2, 7, 8

[53] Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule, 2025. 2, 7

[54] 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. 2, 8, 1

[55] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 8

[56] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. In *NeurIPS*, 2024. 8