One-Minute Video Generation with Test-Time Training

[4]Karan Dalal [2]Daniel Koceja [1,3]Jiarui Xu [5]Yue Zhao [1]Shihao Han
[1]Ka Chun Cheung [1]Jan Kautz [1]Yejin Choi [1,2]Yu Sun [1,3]Xiaolong Wang

[1]NVIDIA [2]Stanford University [3]UC San Diego [4]UC Berkeley [5]UT Austin

CVPR Nashville JUNE 11-15, 2025

Motivation

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.

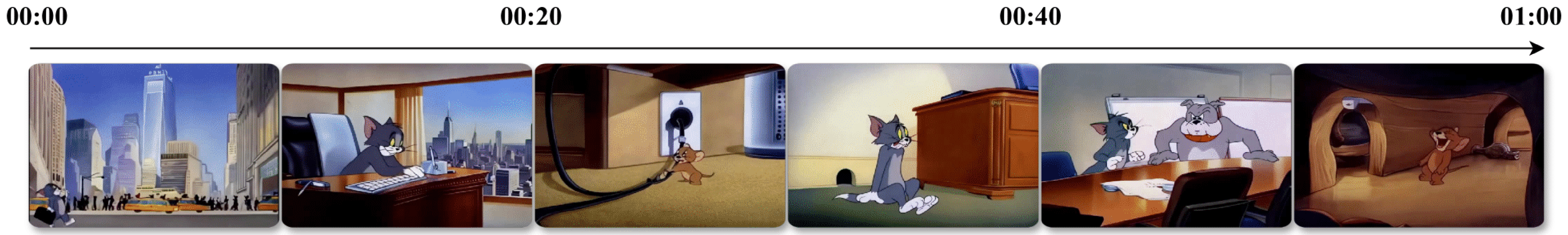
We experiment with Test-Time Training (TTT) layers, whose hidden states themselves can be neural networks, therefore larger and more expressive.

TTT Layers

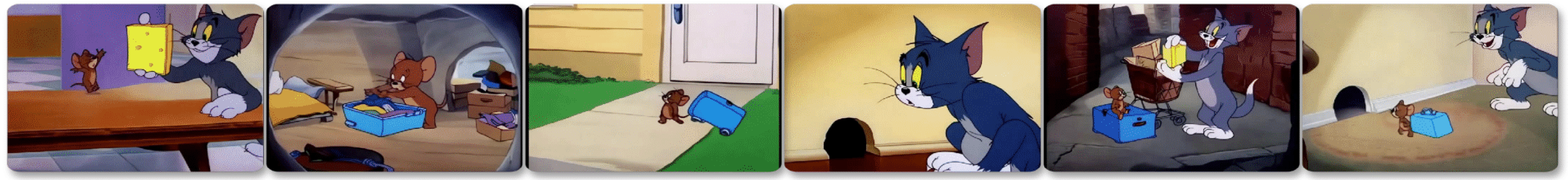
$$\begin{tikzcd}
\text{Output tokens} & & z_1 & & z_{t-1} & z_t = f(x_t; W_t) & \text{Output rule} \\
\text{Hidden state} & W_0 \arrow[r] & W_1 \arrow[r] \arrow[u] & \ldots \arrow[r] & W_{t-1} \arrow[r] \arrow[u] & W_t = W_{t-1} - \eta\nabla\ell(W_{t-1}; x_t) \arrow[u] & \\
 & & & & & & \text{Update rule} \\
\text{Input tokens} & & x_1 \arrow[uu] & & x_{t-1} \arrow[uu] & x_t \arrow[uu] &
\end{tikzcd}$$

All RNN layers can be expressed as a hidden state that transitions according to an update rule. The key idea in [Sun et. al. 2024] is to **make the hidden state itself a model** with weights, and the **update rule a gradient step** on the self-supervised loss.

Updating the hidden state on a test sequence is equivalent to training the model at test time. This process, known as test-time training (TTT), is programmed into TTT layers.



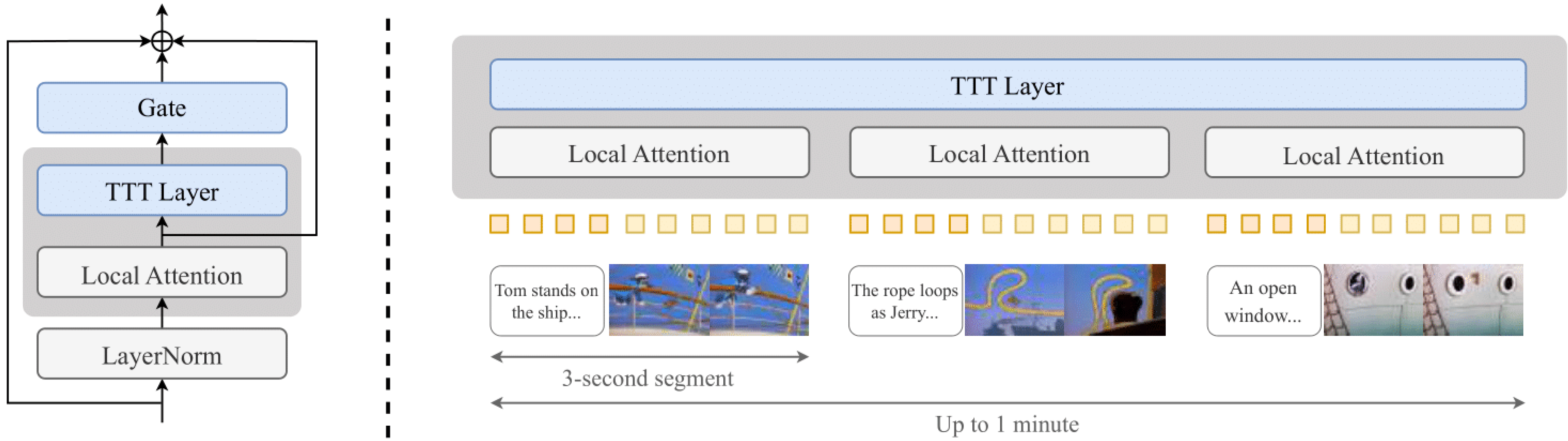
On a sunny morning in New York, Tom, a blue-gray cat carrying a briefcase, arrives at his office in the World Trade Center. As he settles in, his computer suddenly shuts down – Jerry, a mischievous brown mouse, has chewed the cable. A chase ensues, ending with Tom crashing into the wall as Jerry escapes into his mousehole. Determined, Tom bursts through an office door, accidentally interrupting a meeting led by Spike, an irritated bulldog, who angrily sends him away. Safe in his cozy mousehole, Jerry laughs at the chaos.



Jerry happily eats cheese in a tidy kitchen until Tom playfully takes it away, teasing him. Annoyed, Jerry packs his belongings and leaves home, dragging a small suitcase behind him. Later, Tom notices Jerry's absence, feels sad, and follows Jerry's tiny footprints all the way to San Francisco. Jerry sits disheartened in an alleyway, where Tom finds him, gently offering cheese as an apology. Jerry forgives Tom, accepts the cheese, and the two return home together, their friendship restored.

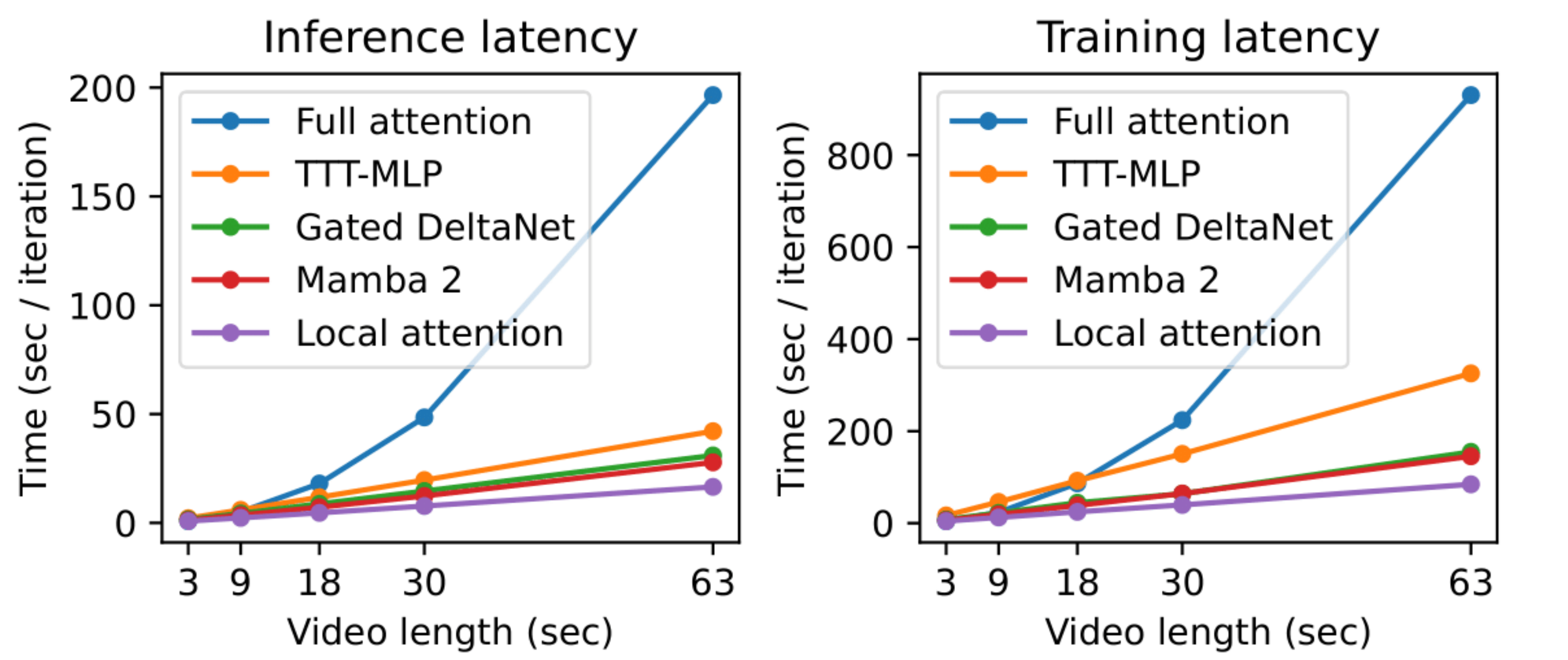
Approach

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. Every video is produced directly by the model in a single shot, **without editing, stitching, or post-processing.**



Training & Inference Efficiency

For 63-second videos, inference with full attention would have taken **11x** longer than local attention, and training **12x** longer. TTT-MLP takes **2.5x** and **3.8x** respectively.



Human Evaluation

TTT layers generate much more coherent videos that tell complex stories, leading by **34 Elo points.**

	Text following	Motion naturalness	Aesthetics	Temporal 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**

Project Website

Additional demo videos and code are available at: