FrameDiT: Diffusion Transformer with Matrix Attention for Efficient Video Generation
Supplementary Materials

Minh Khoa Le[1] Kien Do[2] Duc Thanh Nguyen[3] Truyen Tran [1]
[1] Applied Artificial Intelligence Initiative, Deakin University, Australia
[2] FPT Smart Cloud, Vietnam [3] Deakin University, Australia
[1,3] *{minh.le, duc.nguyen, truyen.tran}@deakin.edu.au*
[2] *kiendd6@fpt.com*

1. Theoretical Proof of Matrix Attention

This section derives the attention maps of our Matrix Attention and compare it with Full 3D and Local Factorized, showing that Local Factorized is our special case. Let $z = [z_t]_{t=1}^{T}$ denote video features, where the feature of frame t is $z_t \in \mathbb{R}^{N\times D}$. We represent index of token as a pair (t, n) where t denotes the frame index, n denotes the spatial token index within the frame. After flattening all tokens in temporal-spatial order, the full attention matrix is $A \in \mathbb{R}^{(TN)\times(TN)}$, with each element $A\left[(t, n), (t', n')\right]$ specifies how much token (t', n') attends to token (t, n). We omit softmax, scaling terms, and MLP biases since they do not affect structure of the attention map.

Full 3D Attention computes output as

$$y = A_{\text{full}} z W_v, \tag{1}$$

where y is the output, attention map A_{full} is unconstrained. As a result, each token (t, n) can directly attend to any token (t', n').

Local Factorized Attention replaces this with spatial and temporal attention. For each frame t, the spatial attention operates only within that frame

$$x_{t,n} = \sum_{n'} S_t\left[n, n'\right] z_{t,n'} W'_v, \tag{2}$$

where $x_{t,n}$ is the output of spatial attention at position (t, n), $S_t \in \mathbb{R}^{N\times N}$ is the spatial attention map. Stacking all frames:

$$x = SzW'_v, \tag{3}$$

with $S = \text{diag}\left(S_1, \ldots, S_T\right) \in \mathbb{R}^{(TN)\times(TN)}$ is the block diagonal matrix, meaning no information exchange across frames. Then, temporal attention compute output for each spatial index n as

$$y_{t,n} = \sum_{t'} H_n\left[t, t'\right] x_{t',n} W''_v, \tag{4}$$

$H_n \in \mathbb{R}^{T\times T}$ is the temporal attention map. We can also rewrite for all spatial indexes in the matrix form:

$$y = xW_q W_k^T x^T x W''_s = HxW''_v, \tag{5}$$

where $H \in \mathbb{R}^{(TN)\times(TN)}$ is attention matrix, defined by

$$\begin{cases} H\left[(t, n), (t', n')\right] = H_n\left[t, t'\right] & \text{if } n = n' \\ H\left[(t, n), (t', n')\right] = 0 & \text{otherwise.} \end{cases} \tag{6}$$

Replace x in Equation (5) by Equation (3):

$$y = HSzW'_v W''_v = A_{\text{fact}} z W_v, \tag{7}$$

which reveals that Local Factorized Attention implicitly assumes attention matrix must be factorized into $A_{\text{fact}} = HS$. Each element of A_{fact} is computed as

$$A_{\text{fact}}\left[(t, n), (t', n')\right] \tag{8}$$

$$= \sum_{i,j} H\left[(t, n), (i, j)\right] S\left[(i, j), (t', n')\right] \tag{9}$$

$$= H\left[(t, n), (t', n)\right] S\left[(t', n), (t', n')\right], \tag{10}$$

since $H\left[(t, n), (i, j)\right] = 0$ if $j \neq n$, and $S\left[(i, j), (t', n')\right] = 0$ if $i \neq n'$. Thus, the interaction between tokens (t, n) and (t', n') relies solely on the single intermediate token (t', n).

$$(t', n') \xrightarrow[\text{spatial}]{S} (t', n) \xrightarrow[\text{temporal}]{H} (t, n).$$

In the FrameDiT-G model, spatial attention remains the same $x = SzW'_v$, while temporal attention is replaced by Matrix Attention, which is computed as

$$y = \underbrace{U_q^\intercal x W_q}_{\text{key}} \underbrace{W_k^T x U_k}_{\text{query}} \underbrace{U_v^\intercal x W''_v}_{\text{value}} \tag{11}$$

$$= U_q^\intercal H U_k U_v^\intercal S z W'_v W''_v. \tag{12}$$

	FVD↓				FVMD↓				FID↓			
	16	32	64	128	16	32	64	128	16	32	64	128
FrameDiT-H-Concat	**66.15**	**126.90**	**225.39**	**256.40**	**943.32**	**478.14**	**125.81**	**70.10**	13.45	15.48	17.54	22.29
FrameDiT-H-Gated	67.55	130.88	233.27	265.25	964.13	528.42	153.65	89.23	**13.14**	**15.29**	**17.50**	**21.42**

Table 1. **Comparison between different Fusion layers for FrameDiT-H**. FrameDiT-H-Concat is the model using Concat+MLP fusion, and FrameDiT-H-Gated is the model using sigmoid fusion described in 2.3.

Letting $H' = U_q^\intercal H U_k U_v^\intercal$, this simplifies to

$$y = H' S z W_v = A_{\text{mat}} z W_v. \tag{13}$$

One noteworthy point is that $H^{'}$ is not constrained like in Equation (6). Each element of A_{mat} is computed as

$$A_{\text{mat}}\left[(t,n),(t',n')\right] \tag{14}$$

$$= \sum_{i,j} H'\left[(t,n),(i,j)\right] S\left[(i,j),(t',n')\right] \tag{15}$$

$$= \sum_{j=1}^{N} H'\left[(t,n),(t',j)\right] S\left[(t',j),(t',n')\right]. \tag{16}$$

Unlike Local Factorized, the interaction between (t,n) and (t',n') may pass through all spatial tokens $\{t',j\}_{j=1}^{N}$.

$$(t',n') \xrightarrow[\text{spatial}]{S} \{(t',j)\}_{j=1}^{n} \xrightarrow[\text{temporal}]{H} (t,n)$$

This removes the single-token bottleneck imposed by Local Factorized Attention while remaining significantly more efficient than Full 3D Attention. In case when $U_q = U_k = U_v = I$, H' reduces to H, and Matrix Attention collapses exactly to the temporal attention used in Local Factorized Attention. Thus, Local Factorized Attention naturally emerges as a special case of our formulation, corresponding to the identity row-weight matrices of Matrix Attention.

2. Additional Experiment

2.1. Additional Implementation Details

We train all models using the DDPM framework with 1000 training diffusion steps. According to [4], we choose to learn both mean and variance of the reverse process to improve log-likelihood, increasing sample quality using the loss function

$$L = L_{\text{simple}} + \lambda L_{\text{vlb}}, \tag{17}$$

with $\lambda = 10^{-3}$.

We employ the noise-parameterization for all models. For the latent representation, we use the Stable Diffusion 2.0 autoencoder [5], encoding each frame independently with a compression ratio of $\{1, 8\}$. We use this VAE to ensure that improvement is solely from the diffusion modeling rather than the reconstruction quality of the autoencoder.

For dataset, we preprocess them by center crop, and resize to desire resolution. We randomly sample with fixed frame interval 3 from videos, except interval 1 for Taichi-HD 128×128 at 128 frames. Our model and all baselines are implemented using the Latte codebase and trained under the exact same settings to ensure fair comparison. All conditioning inputs, including noise levels and class labels, are injected into the network through AdaLN-Zero layers. For all experiments, training is performed in FP16 precision for computational efficiency. We clip maximum norm of gradients to 1.0 from 100,000 training steps to stabilize training, and exponential moving average (EMA) of the model weights with a 0.999 decay. To improve high-resolution quality at 256×256, we jointly train on both videos and images, with 8 images for each video. We use the AdamW optimizer with a learning rate of 1×10^{-4}. We use DDPM sampler with 250 sampling steps for all models. For UCF101, we apply classifier-free guidance with CFG = 7.0.

We evaluate models using both video-level and frame-level metrics. For video-level evaluation, we adopt FVD [6] and FVMD [3]. FVD computes the Fréchet distance between feature distributions of real and generated videos, where features are extracted using a pretrained I3D network [1]; it reflects both overall video quality and temporal coherence. FVMD focuses specifically on motion consistency: it tracks keypoints using a pretrained PIPs++ model [7] to obtain motion trajectories, then measures the Fréchet distance between the real and generated motion features. Since FVMD processes videos in 16-frame chunks, longer sequences lead to more chunks, therefore, produce lower FVMD values. For this reason, we report the relative FVMD improvement over Local Factorized Attention to better reflect motion consistency gains across different sequence lengths. For frame-wise metrics, we report FID [2], evaluating the similarity between real and generated frame distributions using Inception features, providing a measure of overall image realism.

2.2. Additional Result

Comparison between different attention mechanisms Figure 1 present qualitative results of different attention mechanism on Taichi-HD at 128×128 resolution for video lengths of 128 frames. Local Factorized Attention fails to preserve human structure, producing inconsistent body movements; these issues worsen as sequence length increases. In contrast, both Full 3D Attention, FrameDiT-G and FrameDiT-H maintain stable temporal dynamics across all lengths, producing smooth, consistent human motion even for 128-frame videos.

Because the Stable Diffusion 2.0 autoencoder is trained primarily on high-resolution datasets, encoding images at 128×128 may degrade small or fine-grained details such as facial features and hands. As a result, the generated videos may exhibit blurred or distorted small structures, reflecting a limitation of the underlying Image VAE.

Comparision with existing video generative models Figure 2 presents qualitative comparisons on UCF101 dataset between prior video generative models and our approach. Latte often produces temporally inconsistent videos; for example, in the second sample of Latte, the scene begins with a single person but unexpectedly introduces an additional person mid-sequence, revealing weak long-range temporal modeling. This issue comes from its Local Factorized Attention, which cannot preserve frame-level information during temporal attention.

AR-Diffusion produces sharp frames but exhibits minimal motion, partly due to its use of a frame interval of 1 and independent noise levels during training. In contrast, our method produces videos with stable temporal dynamics and maintains coherence even for fast-moving foreground objects, demonstrating stronger modeling of complex motion and global temporal structure.

2.3. Aditional ablation study

Fusion mechanism For FrameDiT-H, we investigate alternative fusion mechanisms that combine local and global temporal features. In addition to the default Concat+MLP fusion, we evaluate a gated fusion variant where a learnable scalar gate adaptively controls the contribution of each branch:

$$e = \sigma(\alpha) \times e_{\text{local}} + (1 - \sigma(\alpha)) \times e_{\text{global}}, \qquad (18)$$

where $\sigma(\cdot)$ denotes the sigmoid function, and α are learnable parameters. We denote the two variants as FrameDiT-H-Concat and FrameDiT-H-Gated, respectively. As shown in Table 1, FrameDiT-H-Concat achieves stronger results on video-level metrics across all sequence lengths, indicating that preserving the full information from both branches leads to better temporal modeling. In contrast, FrameDiT-H-Gated performs slightly better on FID, suggesting that gated fusion can improve frame-wise spatial quality but at the cost of discarding some temporal information. Overall, the results highlight that both local and global temporal information are essential, and that aggressively filtering one branch via a sigmoid gate may hurt video consistency.

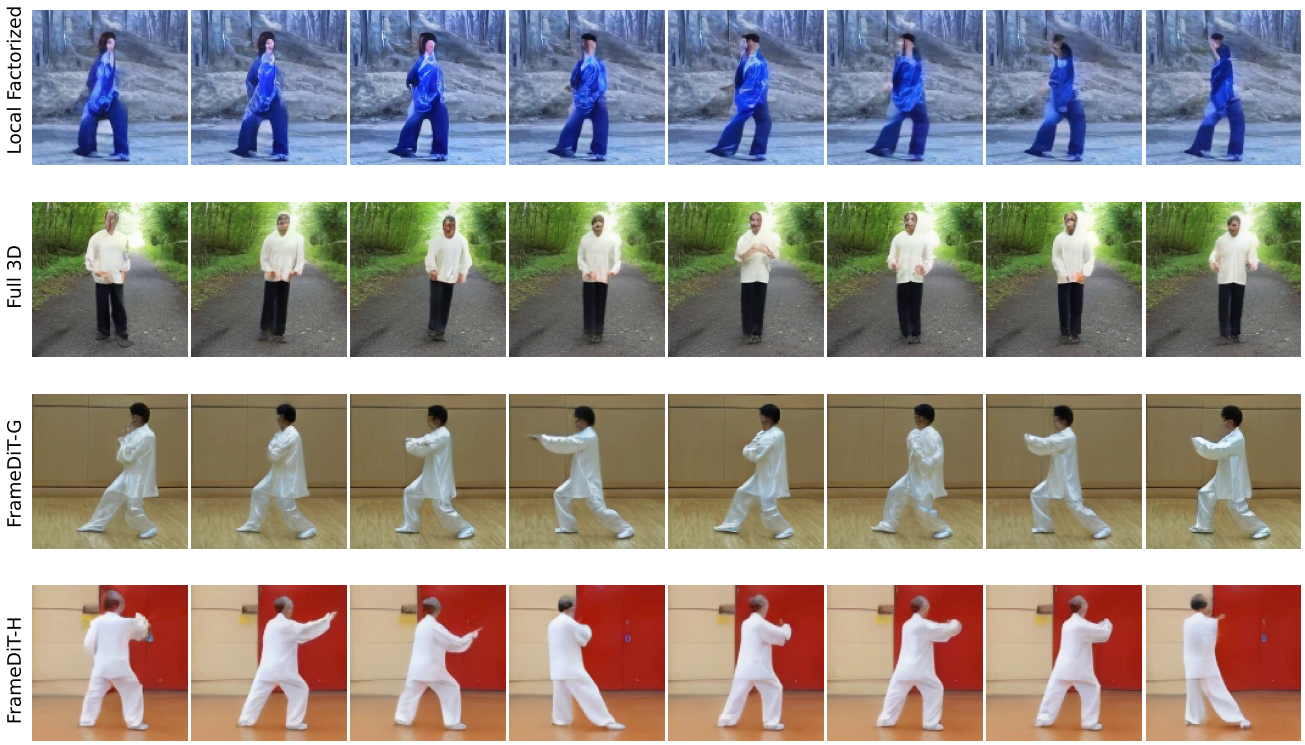

Figure 1. **Qualitative comparison on 128-frame Taichi-HD 128×128.** Local Factorized Attention exhibits severe temporal drift and collapsing human structure. In contrast, Full 3D model and FrameDiT-G, FrameDiT-H remain stable even at 128 frames, generating smooth and coherent motion. The slight blurring of small regions (hands, face) arises from the low-resolution encoding of the Stable Diffusion 2.0 autoencoder.

Figure 2. **Qualitative comparison on UCF101 between prior video generative models and our approach.**

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[3] Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fr\'echet video motion distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124*, 2024. 2

[4] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[6] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 2

[7] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023. 2