Qualcomm

ReHyAt: Recurrent Hybrid Attention for Video Diffusion Transformers

Mohsen Ghafoorian

Amir Habibian

Qualcomm AI Research*

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Introduction

You can turn linear causal attention into RNN!

Sadly, linear attention is under-expressive! :(

Video Diffusion's Bottleneck

- DiTs dominate SOTA video diffusion models
- Quadratic self-attention (Softmax) complexity is the real bottleneck.

Self Attn. Cross Attn. FFN

$$\mathcal{O}(sb(nd^2 + n^2 + nn_k d + (n+n_k)d^2 + nd^2))$$

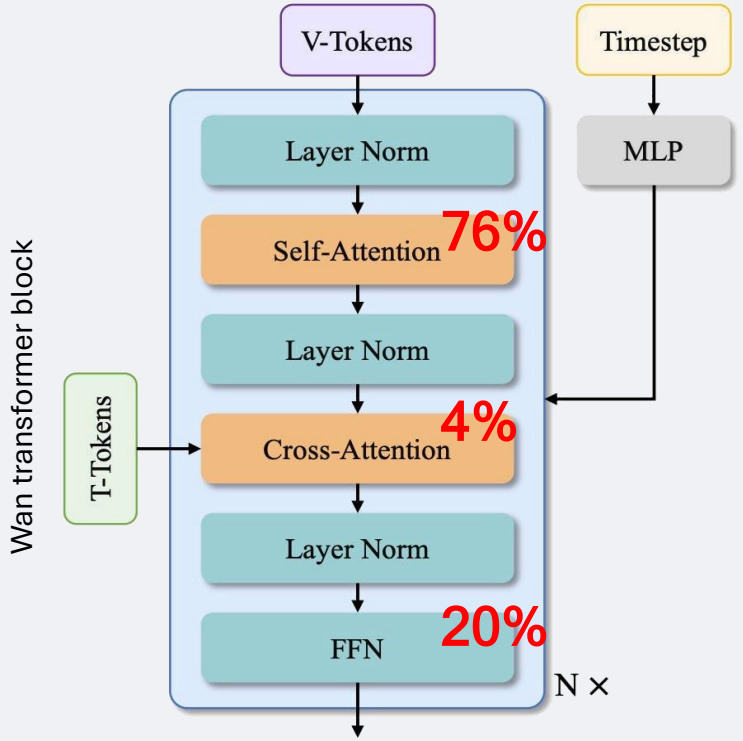
- n: #tokens, s: #denoising steps, d: feature dim, n_k: #context tokens, b: #blocks
- For Wan2.1 1.3B 480p n=~33k

- Linear attention[2]

$$V_i' = \frac{\sum_{j=1}^{N} \text{sim}(Q_i, K_j) V_j}{\sum_{j=1}^{N} \text{sim}(Q_i, K_j)}. \quad \text{sim}(q,k) = \exp\left(\frac{q^T k}{\sqrt{D}}\right)$$

$$V_i' = \frac{\sum_{j=1}^{N} \phi(Q_i)^T \phi(K_j) V_j}{\sum_{j=1}^{N} \phi(Q_i)^T \phi(K_j)}, \qquad V_i' = \frac{\phi(Q_i)^T \sum_{j=1}^{N} \phi(K_j) V_j^T}{\phi(Q_i)^T \sum_{j=1}^{N} \phi(K_j)}.$$

Terms that can be precomputed



Wan[1]2.1 1.3B 480p compute distribution

1. Team Wan et al. "Wan: Open and advanced large-scale video generative models." arXiv preprint arXiv:2503.20314 (2025).
2. Katharopoulos et al. "Transformers are RNNs: Fast autoregressive transformers with linear attention." ICML 2020.

Recurrent Hybrid Attention (ReHyAt)

High-levels

- What?
  - Existing hybrid attention video model, attention surgery*, saves compute, but:
    - Memory and compute still grow quadratically with sequence length.
    - Unable to turn into an RNN.
  - Goal: making a SOTA pretrained VDM (e.g. Wan) an RNN
- How?
  - Define a causal temporally-arranged Linear/Softmax hybrid attention.
  - Divide inference into several temporal chunks; Model the dependencies
    - **within-chunk with Softmax**.
    - to farther **past tokens with the efficient linear attention**
- Outcome?
  - Achieving **VBench of 84.1** for the Hybrid adapted model, within only **<160 GPU-hours**.
  - **Device-friendly RNN** model, that computationally enables long video generation.

Wan2.1 1.3B, VBench: 83.1

ReHyAt, VBench: 84.1

* Ghafoorian et al. “Attention Surgery: An Efficient Recipe to Linearize Your Video Diffusion Transformer." *CVPR* 2026.

Recurrent Hybrid Attention (ReHyAt)

Methods

$$\hat{y}_t = \frac{a_t^{\mathrm{S}} + a_t^{\mathrm{L}}}{n_t^{\mathrm{S}} + n_t^{\mathrm{L}}},$$

$$a_t^{\mathrm{S}} = \sum_{j \in \mathcal{T}_t^{\mathrm{S}}} \exp(Q_t k_j^\top / \sqrt{D} - c_t) v_j,$$

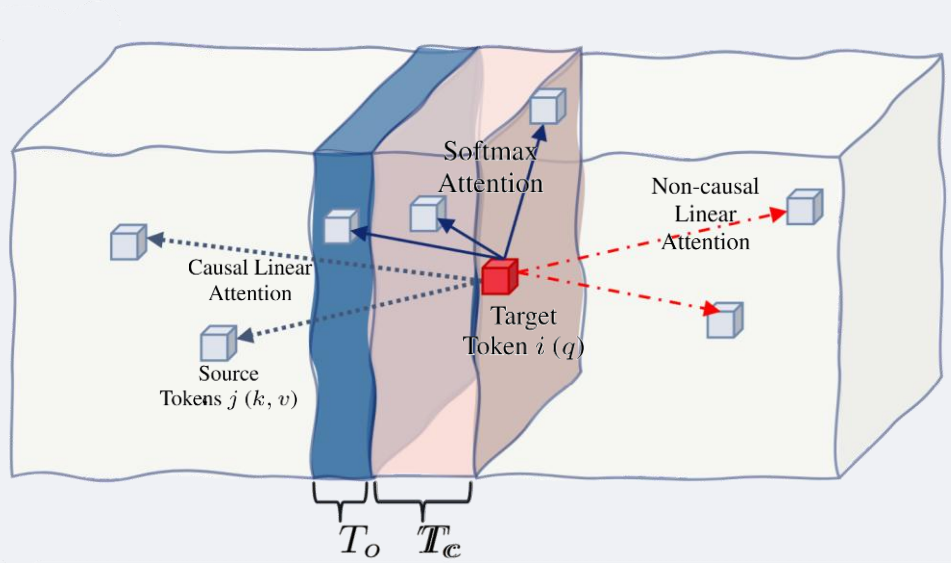
$$a_t^{\mathrm{L}} = \phi_q(Q_t) \Big(\sum_{j \in \mathcal{T}_t^{\mathrm{L}}} \phi_k(k_j)\, v_j^\top \Big),$$

$$n_t^{\mathrm{S}} = \sum_{j \in \mathcal{T}_t^{\mathrm{S}}} \exp(Q_t k_j^\top / \sqrt{D} - c_t),$$

$$n_t^{\mathrm{L}} = \phi_q(Q_t) \Big(\sum \phi_k(k_j) \Big),$$

$$\mathcal{T}_t^{\mathrm{L}} = \{ j \mid j < \max(tN' - T_o HW, 0) \}$$

$$\mathcal{T}_t^{\mathrm{S}} = \{ j \mid \max(tN' - T_o HW, 0) \le j < (t+1)N' \}$$



- RNN formulation:

$$\mathcal{T'}_t^{\mathrm{S}} = \{ j \mid \max(tN' - T_o HW, 0) \le j < (t+1)N' - T_o HW \}.$$

$$s_0 = 0,$$
$$z_0 = 0,$$

Chunk-wise compute bounded by constant $T_c + T_o \ll T$

$$y_t = \frac{\sum_{j \in \mathcal{T}_t^{\mathrm{S}}} \exp(Q_t k_j^\top / \sqrt{D} - c_t) v_j + \phi_q(Q_t)\, s_t}{\sum_{j \in \mathcal{T}_t^{\mathrm{S}}} \exp(Q_t k_j^\top / \sqrt{D} - c_t) + \phi_q(Q_t)\, z_t},$$

$$s_{t+1} = s_t + \sum_{j \in \mathcal{T'}_t^{\mathrm{S}}} \phi_k(k_j)\, v_j^\top,$$

$$z_{t+1} = z_t + \sum_{j \in \mathcal{T'}_t^{\mathrm{S}}} \phi_k(k_j),$$

State Variables Summarizing the past

Memory remains constant & compute grows linear with growth of n.

Linear Attention Feature Mapping And Overview

- Design:
  - d-layer MLP, with degree-P polynomial:

$$\phi(x) = [\ (\psi_1(x))^1, (\psi_2(x))^2, \ldots, (\psi_P(x))^P\ ]^\top \in \mathbb{R}^{D'},$$

  - ReLU non-linearity on the last layer
    - To guarantee the non-negativity requirement of the kernel trick.
- Rationale
  - **Learnable**
    - Opposed to non-learnable elu(x)+1 in original attention
    - Increase the rank and expressiveness of linear branch
  - **Polynomial**:
    - Facilitate approximating the large dynamic range of exponential:

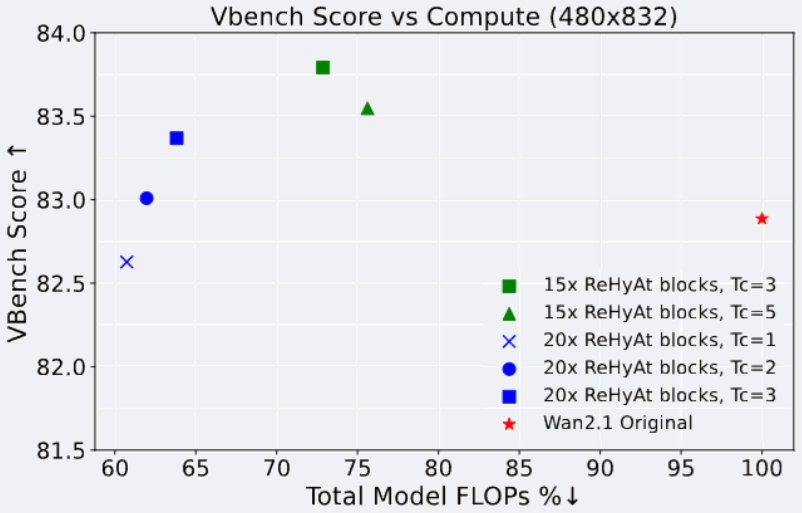
$$e^{q_i k_j^\top} \approx \phi(q_i)\phi(k_j)$$

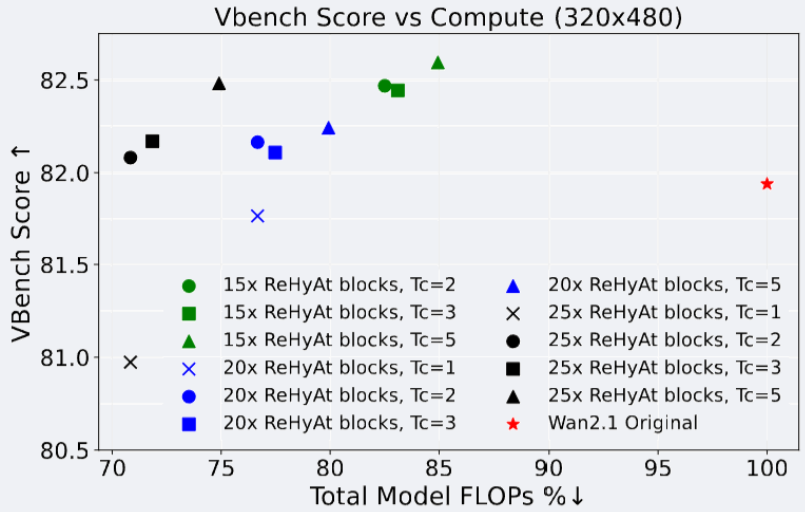
Method Overview:

1. **Block-wise Distillation** of the full bidirectional Softmax attention teacher into the student causal recurrent hybrid attention to pretrain Φ params.
2. **Light-weight End-to-End Finetuning**
3. **Reformulation of the Inference Graph into a Chunk-wise RNN.**

ReHyAt - Quantitative Results

Quality/Compute trade-off vs Wan2.1





Comparison to SOTA video Diffusion

Models with 2B–5B parameters	Total↑	Quality↑	Semantic↑
Open-Sora Plan V1.3 [26]	77.23	80.14	65.62
CogVideoX 5B [44]	81.91	83.05	77.33
CogVideoX1.5 5B [44]	82.01	82.72	79.17
Models up to 2B parameters			
Open-Sora V1.2 [52]	79.76	81.35	73.39
LTX-Video [12]	80.00	82.30	70.79
SnapGenV [41]	81.14	83.47	71.84
Hummingbird 16frame [15]	81.35	83.73	71.84
Mobile Video DiT - Mobile [40]	81.45	83.12	74.76
Mobile Video DiT - Server [40]	83.09	84.65	76.86
CogVideoX 2B [44]	81.55	82.48	77.81
PyramidalFlow [16]	81.72	84.74	69.62
Neodragon [19]	81.61	83.68	73.36
Wan2.1 1.3B [35]	83.31	85.23	75.65
Wan2.1 1.3B* [35]	83.10	85.10	75.12
Linear/Hybrid Models			
Efficient VDiT [8]	76.14	-	-
M4V [13]	81.91	83.36	76.10
STA [50]	83.00	**85.37**	73.52
VSA [49]	82.77	83.60	79.47
SANA-Video [4]	83.71	84.35	**81.35**
Attention Surgery (15×R2) [11]	83.21	85.19	75.25
Wan2.1 1.3B* + ReHyAt (15×T_c=3,T_o=1)	**84.11**	85.03	80.44

VBench

Model	VBench-2.0					
	Total↑	Hum.Fid.↑	Creativity↑	Control.↑	Com.sense↑	Physics↑
Wan2.1 1.3B	56.0	80.7	48.7	**34.0**	63.4	**53.8**
CogVideoX-1.5 5B	53.4	72.1	43.7	29.6	63.2	48.2
Attn. Surgery 15×R2	55.1	78.9	47.5	33.4	63.1	52.8
ReHyAt 15×T_c=3	56.1	**81.9**	55.1	30.8	62.7	50.0
ReHyAt 15×T_c=5	**56.3**	79.8	**55.7**	31.9	**64.2**	49.7

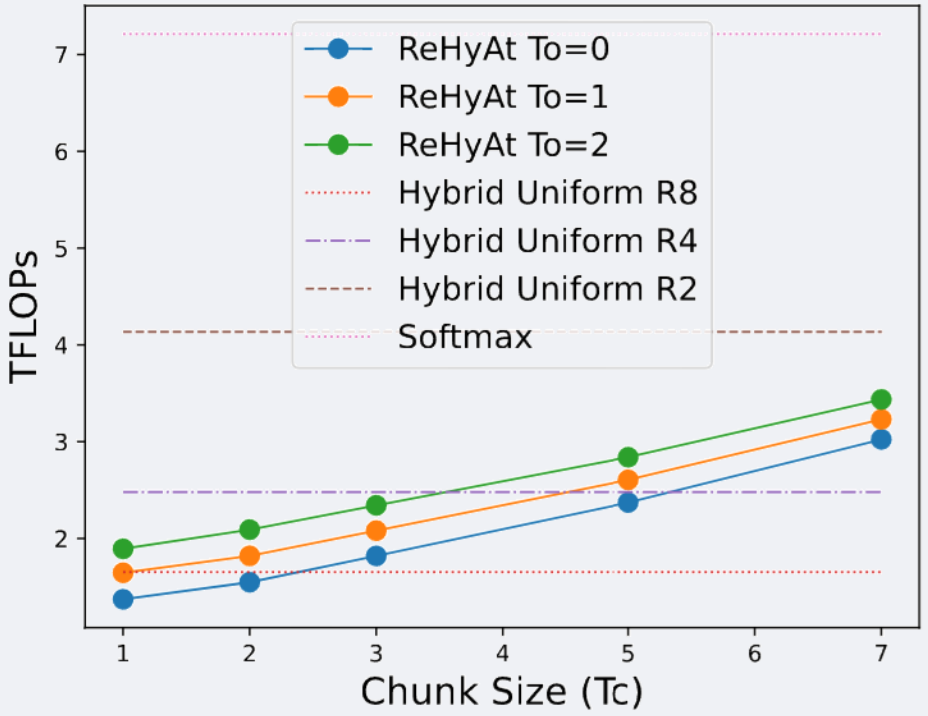
VBench-2.0

Human Preference Study

Prompt Dimension	Human Preference %		
	Ours	No preference	Wan2.1
Color	43.3	46.7	10.0
Human Action	21.7	41.7	36.7
Object Class	25.0	45.0	30.0
Overall Consistency	27.1	47.1	25.9
Scene	40.0	60.0	0.0
Spatial Relationship	20.0	70.0	10.0
Subject Consistency	21.7	28.3	50.0
Temporal Flickering	24.0	54.0	22.0
Temporal Style	43.3	30.0	26.7
Total	27.6	43.5	29.0

Comparison of 15T3O1 to Wan2.1

Ablations



Impact of **chunk size T_c** on 25xReHyAt

Chunk-size T_c	Block TFLOPs↓	VBench		
		Total ↑	Quality ↑	Semantic ↑
1	3.87	80.97	82.37	75.39
2	4.04	82.08	83.86	74.99
3	4.30	82.17	83.72	**75.96**
5	4.82	**82.48**	**84.12**	75.93

Impact of **chunk overlap T_o** on 25xReHyAt

Chunk-overlap T_o	VBench			
	Total ↑	Quality ↑	Semantic ↑	Subj. Cons.↑
0	81.56	83.23	74.90	90.90
1	82.17	83.72	**75.96**	92.05
2	82.17	83.84	75.50	92.13
3	**82.19**	**83.86**	75.51	**92.24**

Impact of **Causal distillation** on 25xReHyAt

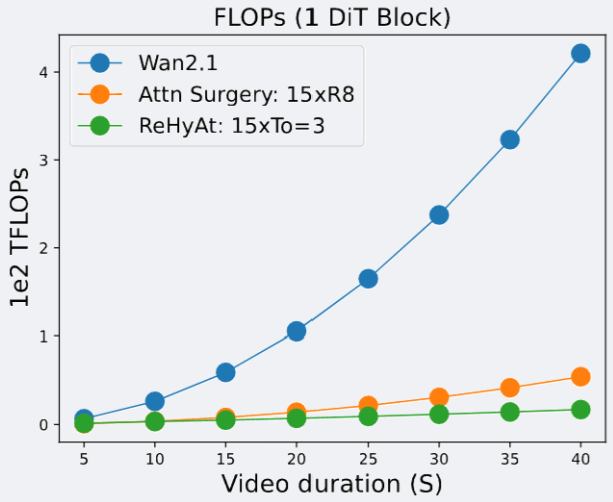
Causal	Block TFLOPs↓	VBench		
		Total ↑	Quality ↑	Semantic ↑
×	4.17	82.27	83.84	**75.99**
✓	**4.04**	**82.35**	**83.97**	75.87

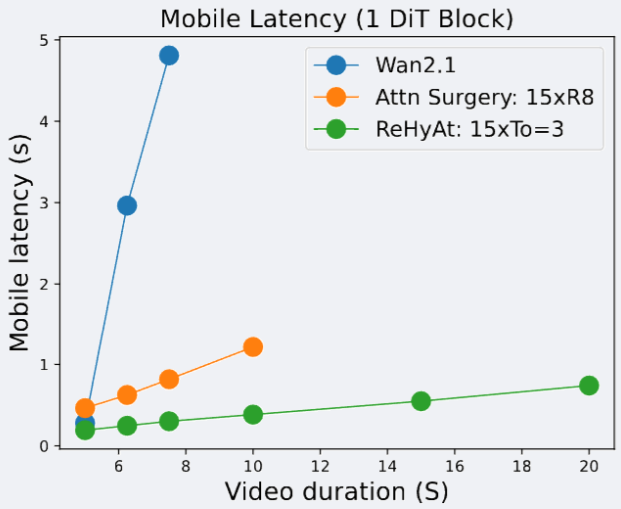
Comparison to T**oken merging** methods

Model (Wan2.1 1.3B)+	ReHyAt	ToMe	USV
VBench Total	84.1	77.1	80.7

Bolya et al. "Token Merging: Your ViT But Faster.“, ICLR 2023

Scaling Behavior and On-device Measurements





Attention Block	Number of frames (320×480) resolution				
	81	101	121	141	161
Softmax Flash Attention	281	2964	4809	OOM	OOM
HedgeHog Linear Attention	360	455	469	542	OOM
Uniform Hybrid - R8	464	625	818	1215	OOM
ReHyAt - T_c=3 (ours)	**192**	**247**	**302**	**329**	**384**

Latency vs number of frames for 1 DiT Block

Attention Block	Number of frames - Memory Read/Write (GB)									
	81		101		121		141		161	
	W	R	W	R	W	R	W	R	W	R
Softmax Flash Attention	5.1	6.0	12.9	16.4	22.7	53.6	OOM	OOM	OOM	OOM
HedgeHog Linear Attention	5.7	8.1	7.0	10.1	6.9	11.3	8.0	13.2	OOM	OOM
Uniform Hybrid - R8	6.3	10.1	5.2	10.9	6.4	13.2	7.8	35.2	OOM	OOM
ReHyAt - T_c=3 (ours)	**1.7**	**2.8**	**2.2**	**3.6**	**2.7**	**4.4**	**3.0**	**4.8**	**3.5**	**5.6**

Total Memory Read/Write of 1 DiT Block

* As measured on Snapdragon 8 – Gen 4