HYLIN: MODALITY-AWARE HYBRID ATTENTION FOR EFFICIENT LONG-CONTEXT VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

While large vision-language models scale to ultra-long sequences, the quadratic complexity of softmax attention becomes a severe bottleneck for both multimodal understanding and video generation. Existing inter-layer hybrids retain expensive full-attention layers, uniform intra-layer hybrids ignore modality-specific structure, and sparse video methods often preserve the full KV cache. We argue that efficient attention must combine *global context compression* with *exact token-level retrieval*. Our analysis reveals substantial heterogeneity across modalities, heads, and depth: visual rows are often diffuse and compressible, whereas text rows motivate exact token-level retrieval, and heterogeneous cross-modal mass makes universal replacement unsafe. We therefore propose HYLIN, a **Hy**brid **Lin**ear architecture centered on a modality-routed hybrid attention module, HLWA. In replaceable VLM layers, HLWA routes visual queries through a constant-state global Linear Attention stream and applies causal softmax attention over all available text keys and values, without a fixed window cap, and then restores multimodal row order. Theory-guided calibration and downstream sensitivity identify high-risk layers, with 25% retained as full attention. A four-stage curriculum performs operator alignment, layer selection, vision-language distillation, and supervised fine-tuning. For video DiTs, we instantiate the principle of combining global compression with exact retrieval with a distinct bidirectional operator. HYLIN-7B closely tracks its full-attention teacher on numerous benchmarks while delivering $2.1\times$ prefill and $3.6\times$ decode speedup together with $3.4\times$ memory reduction at 64K tokens; the video-DiT variant accelerates HunyuanVideo-13B by $1.94\times$ with competitive VBench quality.

1 INTRODUCTION

Vision-Language Models (Bai et al., 2025; Wang et al., 2025b; Li et al., 2025a) and video diffusion Transformers (Dao, 2024; Xi et al., 2025) have converged on a Transformer backbone whose self-attention scales quadratically with sequence length. The cost is no longer asymptotic. A single high-resolution image in Qwen2.5-VL (Bai et al., 2025) produces thousands of patches; a short video clip exceeds 256K tokens and 91 GB of memory; video diffusion models manipulate 100K–500K tokens per denoising step. Hour-long understanding and high-resolution generation now hit the same $\mathcal{O}(N^2)$ wall.

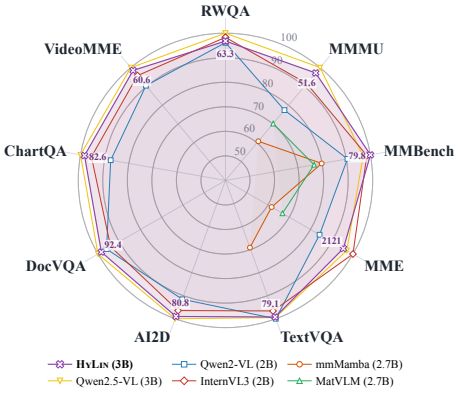


Figure 1: **VLM benchmark comparisons.**

Two main hybrid designs have emerged, but both remain limited. Inter-layer hybrids (Li et al., 2025b; Ren et al., 2025b; Wang et al., 2025c) alternate full- and linear-attention layers, often adopting a 3:1 linear-to-softmax ratio (MiniMax et al., 2025; Zuo et al., 2025; Team et al., 2025), yet still treat all tokens uniformly within each layer. Intra-layer hybrids (Lan et al., 2025; Zhang et al., 2025a; Dong et al., 2025) combine linear and sliding-window attention, but

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

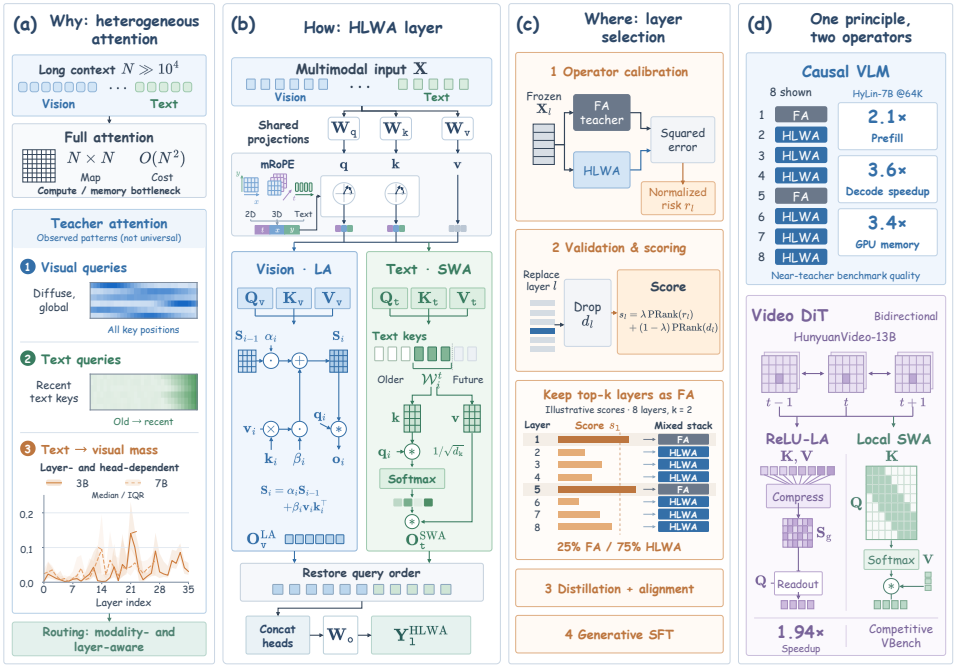


Figure 2: **Overview of HyLin** (a) Heterogeneous visual, textual, and cross-modal attention patterns motivate modality- and layer-aware routing. (b) Shared QKV projections and mRoPE feed visual linear attention and causal text attention, whose outputs are restored to the original query order. (c) Calibrated operator error and single-layer validation drops select the top-k FA layers, followed by distillation and SFT. (d) A causal VLM with 25% FA supports efficient understanding, while a bidirectional DiT combines linear attention with local softmax attention for video generation.

apply both to the same token sequence, overlooking the distinct computational needs of visual and textual tokens. Sparse attention for video generation (Xi et al., 2025; Xu et al., 2025a; Ghafoorian et al., 2026) targets a single domain, still materializes the full KV cache, and does not readily transfer to autoregressive VLMs. To our knowledge, no prior work has validated a shared hybrid-attention design principle across both VLM understanding and video diffusion generation.

Our key insight is that attention should be *modality- and layer-aware*. Our analysis of attention patterns reveals that visual and textual tokens exhibit distinct yet highly layer-dependent behaviors, suggesting that no single attention operator is uniformly optimal across the network. This motivates a selective hybrid design: we apply linear attention where visual context is sufficiently compressible and exact attention over the full causal text prefix for text queries, while retaining full attention in layers that require strong cross-modal interaction or long-range retrieval.

Our analysis yields two findings that challenge uniform hybridization. **(1)** Visual and text queries favor different approximations: diffuse visual attention is amenable to global state compression, whereas text relies on exact token-level retrieval, which we preserve over the full causal text prefix. **(2)** Replaceability varies sharply with depth, so a fixed inter-layer ratio cannot determine where full attention is actually needed. These findings lead to HyLin (**Hy**brid **Lin**ear attention) and its core VLM operator, HLWA (Fig. 2). Within each converted layer, HLWA routes visual queries through Gated DeltaNet (Yang et al., 2025b) and text queries through causal attention over the full text prefix, then restores their original multimodal order. Multimodal RoPE preserves spatial-temporal geometry. Theoretical certificates motivate the row-level risk, while teacher calibration and downstream sensitivity determine which layers retain full attention. A four-stage pipeline then aligns the selected hybrid student with its full-attention teacher. The principle of combining global compression with exact retrieval extends to bidirectional video DiTs through a regime-specific combination of bidirectional ReLU Linear Attention, local attention, and 3D RoPE. As shown in Fig. 2, HyLin-7B approaches its full-attention teacher with **2.1**× prefill and **3.6**× decode speedup together with a **3.4**× memory reduction at 64K tokens, while the Video-DiT variant accelerates HunyuanVideo-13B by **1.94**× with competitive VBench quality.

2 METHOD

Our method is organized around two questions: *how should full attention be replaced when visual and text queries exhibit different structures, and which layers can be converted safely under a fixed compute budget?* We answer the first with our hybrid attention module, HLWA, which combines global visual compression with exact full-prefix text retrieval. We answer the second with calibrated operator risk and downstream sensitivity. Multimodal positional encoding, a regime-specific Video-DiT extension, and a four-stage alignment pipeline complete the design.

2.1 VISUAL LINEAR ATTENTION AND CAUSAL TEXT ATTENTION

Uniform attention replacement assumes that all query rows and layers require the same interaction pattern. Our teacher diagnostics contradict this assumption: visual rows can be diffuse and low-variance, text rows often concentrate on recent text keys, and cross-modal mass varies across heads and depth. This raises our first design question: can a converted layer compress global visual context while preserving exact retrieval over the full text prefix, without running both operators over every token? HLWA answers this by routing visual and text query rows separately. Converted layers perform within-modality mixing, while retained full-attention layers provide direct cross-modal interaction.

Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ contain $N_{\mathtt{v}}$ visual and $N_{\mathtt{t}}$ text tokens, where $N = N_{\mathtt{v}} + N_{\mathtt{t}}$. We use H heads with per-head query/key dimension $d_{\mathtt{k}}$ and value dimension $d_{\mathtt{v}}$, taking $Hd_{\mathtt{v}} = d$. For analysis, fix layer l and head h and suppress both indices. Thus $\mathbf{q}_{\mathtt{i}}, \mathbf{k}_{\mathtt{j}} \in \mathbb{R}^{d_{\mathtt{k}}}$ and $\mathbf{v}_{\mathtt{j}}, \mathbf{o}_{\mathtt{i}} \in \mathbb{R}^{d_{\mathtt{v}}}$. For query i with causal key set $\mathcal{A}_{\mathtt{i}}$, full attention before output projection is

$$z_{\mathtt{ij}} = \frac{\mathbf{q}_{\mathtt{i}}^{\top}\mathbf{k}_{\mathtt{j}}}{\sqrt{d_{\mathtt{k}}}}, \quad p_{\mathtt{ij}} = \frac{\exp(z_{\mathtt{ij}})}{\sum_{\mathtt{r}\in\mathcal{A}_{\mathtt{i}}}\exp(z_{\mathtt{ir}})}, \quad \mathbf{o}_{\mathtt{i}}^{\mathrm{FA}} = \sum_{\mathtt{j}\in\mathcal{A}_{\mathtt{i}}} p_{\mathtt{ij}}\mathbf{v}_{\mathtt{j}}. \tag{1}$$

Global visual compression with Linear Attention. Diffuse visual attention suggests that pairwise retrieval is often redundant in low-risk layers. For the fixed head, we inherit $\mathbf{W}_{\mathtt{q}}, \mathbf{W}_{\mathtt{k}} \in \mathbb{R}^{d\times d_{\mathtt{k}}}$ and $\mathbf{W}_{\mathtt{v}} \in \mathbb{R}^{d\times d_{\mathtt{v}}}$, yielding $\mathbf{q}, \mathbf{k} \in \mathbb{R}^{N\times d_{\mathtt{k}}}$ and $\mathbf{v} \in \mathbb{R}^{N\times d_{\mathtt{v}}}$. Their visual rows satisfy $\mathbf{q}_{\mathtt{v}}, \mathbf{k}_{\mathtt{v}} \in \mathbb{R}^{N_{\mathtt{v}}\times d_{\mathtt{k}}}$ and $\mathbf{v}_{\mathtt{v}} \in \mathbb{R}^{N_{\mathtt{v}}\times d_{\mathtt{v}}}$. HLWA applies

$$\mathbf{O}_{\mathtt{v}}^{\mathrm{LA}} = \mathrm{Linear}\left(\mathbf{q}_{\mathtt{v}}, \mathbf{k}_{\mathtt{v}}, \mathbf{v}_{\mathtt{v}}\right) \in \mathbb{R}^{N_{\mathtt{v}}\times d_{\mathtt{v}}}. \tag{2}$$

Our default VLM mixer is Gated DeltaNet (Yang et al., 2025b); we also evaluate Mamba-2, GLA, and HGRN. These operators summarize the visual sequence in a fixed-size state and require $\mathcal{O}(N_{\mathtt{v}}d_{\mathtt{v}}d_{\mathtt{k}})$ computation rather than pairwise visual attention. Their representative updates and readouts are listed in Appendix Table 5.

We first derive compression diagnostics for a teacher attention row on its full legal key set, then apply them to attention conditioned on visual keys in Appendix C.3. We use a second-order expansion of LogSumExp and Softmax about the row mean. Let $\mathbf{z} = (z_{\mathtt{j}})_{\mathtt{j}=1}^{n} \in \mathbb{R}^n$ collect the logits of one row, with $n = |\mathcal{A}_{\mathtt{i}}|$, $\mathrm{LSE}(\mathbf{z}) = \log\sum_{\mathtt{j}} e^{z_{\mathtt{j}}}$, $\bar{z} = n^{-1}\sum_{\mathtt{j}} z_{\mathtt{j}}$, and $\delta_{\mathtt{j}} = z_{\mathtt{j}} - \bar{z}$. A second-order expansion around the row mean gives

$$J(\mathbf{z}) := \mathrm{LSE}(\mathbf{z}) - \bar{z} - \log n = \log\left(\frac{1}{n}\sum_{\mathtt{j}=1}^{n} e^{\delta_{\mathtt{j}}}\right) = \frac{1}{2}\,\mathrm{Var}_{\mathbf{u}}(\mathbf{z}) + \mathcal{O}(\|\boldsymbol{\delta}\|_{\infty}^{3}), \tag{3}$$

Restoring the row index and using $\nabla_{\mathbf{z}}\,\mathrm{LSE}(\mathbf{z}) = \mathrm{Softmax}(\mathbf{z})$, the same expansion gives the first-order approximation $p_{\mathtt{ij}}^{(1)} = (1+\delta_{\mathtt{ij}})/n_{\mathtt{i}}$ and $\mathbf{o}_{\mathtt{i}}^{(1)} = \sum_{\mathtt{j}\in\mathcal{A}_{\mathtt{i}}} p_{\mathtt{ij}}^{(1)}\mathbf{v}_{\mathtt{j}}$.

Proposition 1 (Centered-Softmax linearization)**.** *Let $\rho_{\mathtt{i}} = \max_{\mathtt{j}\in\mathcal{A}_{\mathtt{i}}}|\delta_{\mathtt{ij}}|$ be the centered-logit radius. Let $D_{\mathtt{i}} = \mathrm{diam}\{\mathbf{v}_{\mathtt{j}} : \mathtt{j}\in\mathcal{A}_{\mathtt{i}}\}$ be the value diameter. Then*

$$\left\|\mathbf{o}_{\mathtt{i}}^{\mathrm{FA}} - \mathbf{o}_{\mathtt{i}}^{(1)}\right\|_2 \le \frac{D_{\mathtt{i}}}{2}\rho_{\mathtt{i}}^{2}. \tag{4}$$

Appendix C gives the centered key–value state form of $\mathbf{o}_{\mathtt{i}}^{(1)}$ and the proof. This is a local compression diagnostic rather than a guarantee for a particular learned LA operator, so every candidate is still validated against measured LA error. A complementary certificate beyond the local Taylor regime follows from the entropy deficit.

(a) Attention routing and mixing.

Variant	Composition	Avg. QA ↑
LA only	V-LA / T-LA	56.6
SWA only	V-SWA / T-SWA	58.6
Intra-layer	LA+SWA	60.6
Inter-layer	LA / SWA	60.0
Modality-aware (ours)	**V-LA / T-softmax**	**71.8**

(b) Multimodal RoPE.

Variant	Avg. QA ↑	VideoMME ↑
NoPE	71.2	48.5
Text-only RoPE	72.8	49.6
Vision-only RoPE	75.8	52.5
Flattened 1D RoPE	74.0	50.4
Multimodal RoPE	**77.8**	**54.0**

(c) Retained full-attention ratio.

FA layers (ratio)	Avg. QA ↑	VideoMME ↑
4 / 36 (11%)	71.8	49.8
9 / 36 (25%)	**77.8**	**54.0**
18 / 36 (50%)	78.5	54.7
27 / 36 (75%)	79.0	55.2

(d) Visual linear-attention operator.

Variant	Avg. QA ↑	VideoMME ↑
GLA (Yang et al., 2024)	76.8	52.3
Mamba-2 (Dao & Gu, 2024)	77.2	52.9
HGRN (Qin et al., 2023)	76.4	51.9
GDN (Yang et al., 2025b)	**77.8**	**54.0**

Table 1: **Ablation studies of HLWA.** (a) Routing and mixing; (b) positional encoding; (c) retained FA ratio; (d) visual mixer. In the modality-aware variant, text softmax attention covers the full causal text prefix; SWA denotes the sliding-window operators in the comparison variants. Avg. QA averages ChartQA, AI2D, and MMBench. The available training settings are summarized in Appendix F.1.

Proposition 2 (Entropy-deficit certificate for global compression)**.** *Let* $n_{\mathtt{i}} = |\mathcal{A}_{\mathtt{i}}|$, $\mathbf{p}_{\mathtt{i}} \in \Delta^{n_{\mathtt{i}}-1}$, $D_{\mathtt{i}} = \max_{\mathtt{j},\mathtt{k}\in\mathcal{A}_{\mathtt{i}}} \|\mathbf{v}_{\mathtt{j}} - \mathbf{v}_{\mathtt{k}}\|_2$, $H(\mathbf{p}_{\mathtt{i}}) = -\sum_{\mathtt{j}} p_{\mathtt{ij}} \log p_{\mathtt{ij}}$, *and* $\Delta_{\mathtt{i}} = \log n_{\mathtt{i}} - H(\mathbf{p}_{\mathtt{i}})$. *If* $\bar{\mathbf{v}}_{\mathtt{i}} = n_{\mathtt{i}}^{-1} \sum_{\mathtt{j}\in\mathcal{A}_{\mathtt{i}}} \mathbf{v}_{\mathtt{j}} \in \mathbb{R}^{d_v}$, *then*

$$\left\|\mathbf{o}_{\mathtt{i}}^{\mathrm{FA}} - \bar{\mathbf{v}}_{\mathtt{i}}\right\|_2 \le D_{\mathtt{i}} \min\left\{1, \sqrt{\frac{\Delta_{\mathtt{i}}}{2}}\right\}. \tag{5}$$

The proof is in Appendix C. These certificates characterize compression on a specified key set; they do not justify deleting cross-modal keys. For a visual query, let $\boldsymbol{\mu}_i^V$ and $\boldsymbol{\mu}_i^T$ be the teacher's conditional value means over legal visual and text keys, and let $\tau_i^{v\to t}$ be its attention mass on text keys. For a nonempty visual key set and the actual visual-only output $\mathbf{o}_i^{\mathrm{LA}}$, the triangle inequality gives

$$\|\mathbf{o}_i^{\mathrm{FA}} - \mathbf{o}_i^{\mathrm{LA}}\|_2 \le \tau_i^{v\to t}\|\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_i^V\|_2 + \|\boldsymbol{\mu}_i^V - \mathbf{o}_i^{\mathrm{LA}}\|_2. \tag{6}$$

The first term measures omitted text access; the second measures visual compression. Appendix C.3 defines the conditional distributions, gives the exact vector decomposition, and reapplies both certificates within the visual key set. Neither certificate bounds the learned mixer directly. We therefore measure the total visual replacement error:

$$\epsilon_{\mathtt{i}}^{\mathrm{LA}} = \left\|\mathbf{o}_{\mathtt{i}}^{\mathrm{LA}} - \mathbf{o}_{\mathtt{i}}^{\mathrm{FA}}\right\|_2, \qquad i \in \mathcal{V}, \tag{7}$$

as the visual component of the layer risk. This quantity already includes both terms of the vector decomposition; no separate omission penalty is added to it.

Text attention. We apply causal softmax attention over the full text prefix. For text query i, the retained key set contains every causally legal text key,

$$\mathcal{P}_{\mathtt{i}}^t = \mathcal{A}_{\mathtt{i}} \cap \mathcal{T}, \qquad n_{\mathtt{i}}^t = |\mathcal{P}_{\mathtt{i}}^t|.$$

Visual keys are excluded from this text-attention operation. Here $n_{\mathtt{i}}^t$ is the number of text keys visible under the causal mask, not a tunable window size. Although attention mass may favor recent text keys, the text branch does not truncate older text keys or impose a fixed window cap. Using the inherited projections and logits from Eq. (1), the text output is

$$\mathbf{o}_{\mathtt{i}}^{\mathrm{text}} = \sum_{\mathtt{j}\in\mathcal{P}_{\mathtt{i}}^t} \frac{\exp(z_{\mathtt{ij}})}{\sum_{\mathtt{r}\in\mathcal{P}_{\mathtt{i}}^t}\exp(z_{\mathtt{ir}})}\mathbf{v}_{\mathtt{j}}, \qquad i \in \mathcal{T}. \tag{8}$$

Stacking these outputs gives $\mathbf{O}_{\mathtt{t}}^{\mathrm{text}} \in \mathbb{R}^{N_{\mathtt{t}}\times d_v}$ per head. At fixed head dimensions, the text branch requires $\mathcal{O}(N_{\mathtt{t}}^2)$ prefill computation and an $\mathcal{O}(N_{\mathtt{t}})$ KV cache. During autoregressive decoding, the text prefix and its KV cache grow as new text tokens are appended; each new query attends to

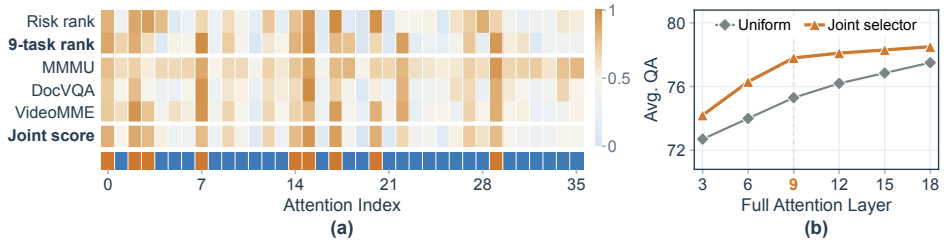


Figure 3: **Layer-selection diagnostics in HYLIN-3B.** (a) Layer-wise signals and Top-9 routing (orange: FA; blue: HLWA). (b) Comparison of Uniform and joint selection at full attention layers. Avg. QA averages ChartQA, AI2D, and MMBench.

the entire available text prefix. At the retained FA layers, the same text token can also attend to causally available visual keys and values. Within HLWA, it does not read the visual recurrent state. The reported HLWA models use this text-attention rule throughout VLM calibration, training, and evaluation.

Modality-wise reassembly. Unlike additive hybrids that evaluate LA and SWA on the same query rows, HLWA assigns exactly one operator to each modality. It restores the visual-LA and text-attention rows to their original query positions and concatenates the H heads. Denoting this operation by Reassemble and using the shared projection $\mathbf{W}_{\mathtt{o}} \in \mathbb{R}^{Hd_{\mathtt{v}} \times d}$, the output is

$$\mathbf{Y}_1^{\mathrm{HLWA}} = \mathrm{Reassemble}\big(\mathbf{O}_{\mathtt{v}}^{\mathrm{LA}}, \mathbf{O}_{\mathtt{t}}^{\mathrm{text}}\big)\, \mathbf{W}_{\mathtt{o}} \in \mathbb{R}^{N \times d}. \tag{9}$$

Retained FA layers introduce visual information into text representations by attending over the legal multimodal context. The residual stream carries these representations through subsequent HLWA layers, which perform visual mixing and text refinement within their respective modalities. Consecutive HLWA layers therefore process visually conditioned text representations without new direct visual retrieval. The default 3B model retains 25% FA layers, selected as described in §2.2.

Multimodal RoPE Linearizing visual interactions should not discard their spatial-temporal geometry. We apply Multimodal RoPE to queries and keys *after* the linear projections and *before* the mixer, and leave values unrotated. Visual coordinates are $(x, y) \in \mathbb{Z}^2$ for image patches and $(t, x, y) \in \mathbb{Z}^3$ for video patches, while text follows the standard mRoPE coordinates. Splitting the $d_{\mathtt{k}}$ rotary dimensions across the video axes gives

$$\mathbf{R}(\mathbf{c}) = \mathrm{diag}\left(\mathbf{R}_{\mathtt{t}}(t), \mathbf{R}_{\mathtt{h}}(x), \mathbf{R}_{\mathtt{w}}(y)\right) \in \mathbb{R}^{d_{\mathtt{k}} \times d_{\mathtt{k}}}, \tag{10}$$

with the temporal block omitted for images. The RoPE identity $\mathbf{R}(\mathbf{c}_{\mathtt{i}})^{\top}\mathbf{R}(\mathbf{c}_{\mathtt{j}}) = \mathbf{R}(\mathbf{c}_{\mathtt{j}} - \mathbf{c}_{\mathtt{i}})$ makes visual logits depend on relative spatial-temporal displacement.

Table 1 examines attention routing, positional encoding, the retained-FA budget, and the visual mixer under the same Qwen2.5-VL-3B initialization, data, and training budget. Table 1a compares routing and mixing strategies with 4 of 36 layers retained as FA (approximately 11%) and the remaining 32 layers converted, using GDN for all LA branches. Intra-layer combines LA and SWA within each converted layer, Inter-layer assigns the two operators to different converted layers, and HLWA routes visual queries to LA and text queries to softmax attention over the full causal text prefix. The retained FA layers provide direct cross-modal interaction in all configurations. Panels (b)–(d) evaluate the hybrid model using the default 25% FA setting except for the factor being varied.

The retained-FA ablations in Table 1 favor Multimodal RoPE over the reduced positional schemes and GDN over the other tested visual mixers. With Multimodal RoPE and GDN, retaining 9 of 36 FA layers under joint selection reaches 77.8 Avg. QA and 54.0 VideoMME. Increasing the FA budget to 18 or 27 layers yields smaller further gains under the same ablation training. These results characterize the positional encoding, visual mixer, and FA budget within the hybrid architecture.

Bidirectional video attention. The causal VLM operator cannot be copied unchanged to Video DiTs: every latent token is visual, and denoising requires bidirectional context. What transfers is the principle of combining global compression with exact retrieval, not the modality split. Here exact retrieval is local: we apply bidirectional ReLU linear attention and sliding-window attention (SWA)

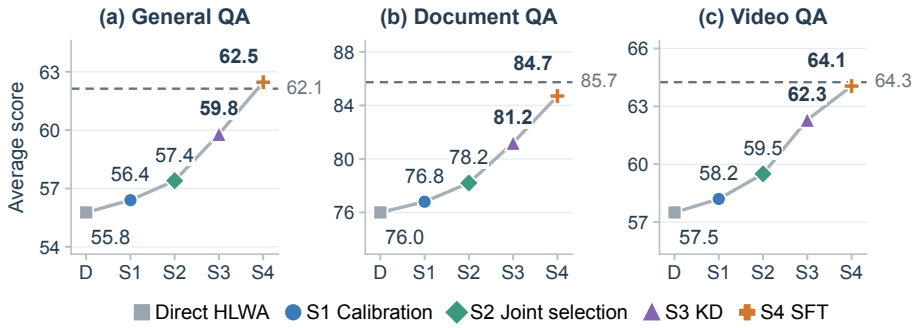


Figure 4: **Progressive recovery across the HyLin pipeline (3B).** (a–c) Average General, Document, and Video QA scores. Dashed lines and their labels mark the FA teacher references. D denotes direct HLWA replacement before calibration.

to all visual latent tokens. For $\mathbf{q}_i, \mathbf{k}_i \in \mathbb{R}^{d_k}$ and $\mathbf{v}_i \in \mathbb{R}^{d_v}$, let $\phi = \mathrm{ReLU}$, $\phi_{R,i}(\mathbf{a}) = \mathbf{R}_{3D}(\mathbf{c}_i)\phi(\mathbf{a})$, $\mathbf{S}_g = \sum_j \phi_{R,j}(\mathbf{k}_j)\mathbf{v}_j^\top \in \mathbb{R}^{d_k \times d_v}$, and $\mathbf{z}_g = \sum_j \phi(\mathbf{k}_j) \in \mathbb{R}^{d_k}$. The global output is

$$\mathbf{o}_i^{\text{bi-LA}} = \frac{\phi_{R,i}(\mathbf{q}_i)^\top \mathbf{S}_g}{\phi(\mathbf{q}_i)^\top \mathbf{z}_g + \varepsilon_{LA}} \in \mathbb{R}^{d_v}. \tag{11}$$

All visual tokens contribute to the bidirectional global state, while SWA preserves exact local interactions. Because the Video-DiT instantiation uses a distinct bidirectional operator, Proposition 3 applies only to the causal VLM. Appendix D.2 discusses the global-branch normalization and summarizes the adaptation, sampling, and evaluation settings.

- **Preserve multimodal geometry.** At the default 25% FA setting, Multimodal RoPE outperforms the tested reduced positional schemes within the HLWA architecture.
- **Combine GDN with retained FA layers.** GDN is the strongest tested visual mixer at 25% retained FA; increasing the FA budget yields smaller further gains under the same ablation training.

2.2 Error-Guided Calibration and Layer Selection

Even a strong replacement operator should not be applied uniformly across depth: the same approximation error can be benign in one layer and damaging in another. This raises our second design question: under a fixed FA budget, which layers require exact attention? We answer it by separating local operator fidelity from downstream sensitivity. Visual replacement error includes both omitted text access and visual compression; the text branch omits direct attention to visual keys while retaining all legal text keys. The total visual-row and text-row errors occur on disjoint rows, so their squared norms add at the layer level. The two components within a visual row need not be orthogonal.

Proposition 3 (HLWA layer-error decomposition)**.** *Fix the teacher input, attention mask, and inherited projections for one layer. For each text query, let* $\mathcal{R}_i = \mathcal{A}_i \setminus \mathcal{P}_i^t$, $\tau_i = \sum_{j \in \mathcal{R}_i} p_{ij}$, *and* $\boldsymbol{\mu}_i^P, \boldsymbol{\mu}_i^R \in \mathbb{R}^{d_v}$ *be the retained and omitted value means. For each head, set* $e_i = u_i = \epsilon_i^{\mathrm{LA}}$ *for* $i \in \mathcal{V}$, *and* $e_i = \tau_i \|\boldsymbol{\mu}_i^R - \boldsymbol{\mu}_i^P\|_2$, $u_i = \tau_i D_i$ *for* $i \in \mathcal{T}$, *with both text terms zero when* $\tau_i = 0$. *Let* $\Delta \mathbf{O}_l = \mathbf{O}_l^{\mathrm{HLWA}} - \mathbf{O}_l^{\mathrm{FA}} \in \mathbb{R}^{N \times H d_v}$. *Then*

$$\mathcal{E}_l := \|\Delta \mathbf{O}_l\|_F^2 = \sum_h \sum_i e_i^2 \le \mathcal{U}_l := \sum_h \sum_i u_i^2. \tag{12}$$

The shared output projection scales this squared error by at most $\|\mathbf{W}_o\|_2^2$.

The proof is in Appendix C. Since the text branch retains all legal text keys, $\mathcal{R}_i = \mathcal{A}_i \cap \mathcal{V}$ contains only omitted visual keys: the text-row error measures cross-modal omission, not truncation of the text prefix. No text-window parameter is selected. Because $\mathcal{U}_l$ still depends on the learned visual-LA error, every candidate is calibrated against the frozen teacher.

Stage 1: Layer-wise output calibration. Each candidate receives the same teacher hidden states $\mathbf{X}_{\mathtt{l}} \in \mathbb{R}^{N\times d}$. We copy the pretrained projections and optimize only the new LA parameters by mean pre-projection MSE. On calibration set $\mathcal{C}$, the normalized realized risk and its calibration-dependent upper bound are

$$r_{\mathtt{l}} := \mathbb{E}_{\mathbf{X}\sim\mathcal{C}}\left[\frac{\mathcal{E}_{\mathtt{l}}}{\|\mathbf{O}_{\mathtt{l}}^{\mathrm{FA}}\|_{\mathtt{F}}^2+\varepsilon_{\mathtt{r}}}\right] \le \mathcal{R}_{\mathtt{l}}^{\mathrm{th}} := \mathbb{E}_{\mathbf{X}\sim\mathcal{C}}\left[\frac{\mathcal{U}_{\mathtt{l}}}{\|\mathbf{O}_{\mathtt{l}}^{\mathrm{FA}}\|_{\mathtt{F}}^2+\varepsilon_{\mathtt{r}}}\right]. \tag{13}$$

The upper bound $\mathcal{R}_{\mathtt{l}}^{\mathrm{th}}$ is semi-empirical: its visual term uses measured replacement error, while its text term uses the attention-mass and value-diameter bound. These quantities describe a single replacement at fixed teacher inputs; they do not guarantee network stability when multiple layers are converted simultaneously.

Stage 2: Full-attention layer selection. Let $\mathcal{M}^{\mathrm{FA}}$ denote the teacher and $\mathcal{M}^{(\mathtt{l})}$ replace only layer l with its calibrated HLWA candidate. We measure downstream sensitivity on nine validation benchmarks $\mathcal{B}_{\mathrm{sel}}$ (including MMMU, DocVQA, and VideoMME) using the relative drops in their higher-is-better scores m_b:

$$\delta_{\mathtt{l}}^{(b)} := \left[\frac{m_b(\mathcal{M}^{\mathrm{FA}})-m_b(\mathcal{M}^{(\mathtt{l})})}{m_b(\mathcal{M}^{\mathrm{FA}})+\varepsilon_{\mathtt{d}}}\right]_+, \qquad d_{\mathtt{l}} := \frac{1}{|\mathcal{B}_{\mathrm{sel}}|}\sum_{b\in\mathcal{B}_{\mathrm{sel}}}\delta_{\mathtt{l}}^{(b)}. \tag{14}$$

where $[x]_+ = \max(x,0)$ clips negative drops to zero, and relative normalization makes metrics with different scales comparable. We combine $d_{\mathtt{l}}$ with the realized operator risk $r_{\mathtt{l}}$ as

$$s_{\mathtt{l}} = \lambda\ \mathrm{PRank}(r_{\mathtt{l}}) + (1-\lambda)\ \mathrm{PRank}(d_{\mathtt{l}}). \tag{15}$$

with $\lambda \in [0,1]$. For $\mathbf{x} \in \{\mathbf{r}, \mathbf{d}\}$, PRank uses ascending ranks across all $L > 1$ candidate layers, assigning average ranks to ties:

$$\mathrm{PRank}(x_{\mathtt{l}}) = \frac{\sum_{\mathtt{j}\neq\mathtt{l}}\mathbf{1}[x_{\mathtt{j}} < x_{\mathtt{l}}] + \frac{1}{2}\sum_{\mathtt{j}\neq\mathtt{l}}\mathbf{1}[x_{\mathtt{j}} = x_{\mathtt{l}}]}{L-1}, \tag{16}$$

where $\mathbf{1}[\cdot]$ is the indicator function. The resulting ranks lie in $[0,1]$, with larger values indicating greater replacement risk. Joint uses $\lambda = 0.6$ and retains the k highest-scoring layers as FA, breaking score ties in favor of smaller layer indices. This empirical selection determines where direct cross-modal interaction remains available, without guaranteeing lossless conversion of the other layers. Figure 3(a) visualizes the selection signals, and Appendix F.1 compares Joint with Uniform, Bound-only, Realized-only, and Benchmark-only under the same 9/36 FA-layer budget and 3M-example direct-SFT protocol. Joint achieves 77.8 Avg. QA and 54.0 VideoMME, exceeding Benchmark-only by 1.3 and 0.9 percentage points, respectively.

Stage 3: Knowledge distillation. The assembled hybrid student is then aligned to the frozen full-attention teacher at the output distribution. Writing $f_{\mathtt{orig}}$ and $f_{\mathtt{hybrid}}$ for the teacher's and student's next-token distributions over target text positions, we optimize

$$\mathcal{L}_{\mathtt{KD}} = D_{\mathtt{KL}}\big(f_{\mathtt{orig}}(\mathbf{X}) \,\|\, f_{\mathtt{hybrid}}(\mathbf{X})\big). \tag{17}$$

Stage 4: Generative finetuning. We then unfreeze both branches and optimize next-token prediction over target text positions. For interleaved layouts, the visual context $\mathbf{H}_{\mathtt{v}}^{(\mathtt{i})}$ includes only visual tokens that are legal under the causal mask for target position i; future visual tokens are never exposed. Writing $\mathbf{H}_{\mathtt{v}}^{(\mathtt{i})}$ for that position-wise legal context,

$$\mathcal{L}_{\mathtt{SFT}} = -\frac{1}{|\mathcal{T}_{\mathtt{target}}|}\sum_{\mathtt{i}\in\mathcal{T}_{\mathtt{target}}}\log P\Big(X_{\mathtt{i}} \mid X_{<\mathtt{i}}, \mathbf{H}_{\mathtt{v}}^{(\mathtt{i})}\Big). \tag{18}$$

Calibration and selection establish the architecture, logit-level distillation restores model-level compatibility, and generative finetuning closes the remaining gap across the benchmark groups. Because Stages 2–4 change both architecture and training, the cumulative recovery shown in Figure 4(a–c) describes the pipeline rather than isolating each module under a matched token budget.

- **Local fidelity is not downstream sensitivity.** Operator error and task degradation capture complementary failure modes; joint Top-k routing spends exact attention where either failure is costly and converts the remainder to HLWA.

Table 2: **Image-understanding results.** Best linear/hybrid results are in **bold**.

Model	# Params	General				Visual QA			Document			
		MMMU	MME	MMB	MMStar	AI2D	ChartQA	TextVQA	DocVQA	InfoVQA	OCRBench	RWQA
Smaller LMMs (≤4B)												
Mono-InternVL-1.5 (Luo et al., 2025)	1.8B	39.1	-	64.0	-	67.4	72.2	73.7	81.7	47.9	801	-
DeepSeek-VL (Lu et al., 2024)	2B	32.2	1532	64.6	-	51.5	-	57.8	-	-	-	-
MiniCPM-V (Yao et al., 2025)	4B	51.2	2167.7	79.1	59.0	80.9	73.0	81.4	94.0	67.0	862	66.0
InternVL3 (Zhu et al., 2025)	2B	48.6	2221.2	78.6	60.7	78.7	80.2	77.0	88.3	66.1	835	64.3
Qwen2-VL (Wang et al., 2024c)	2B	41.1	1872.0	72.2	48.0	74.7	73.5	79.7	90.1	65.5	809	62.9
Qwen2.5-VL (Bai et al., 2025)	3B	53.1	2157.0	77.4	55.9	81.6	84.0	79.3	93.9	77.1	797	65.4
Smaller Linear LMMs (≤4B)												
VL-Mamba (Qiao et al., 2024)	3B	-	1369.6	57.0	-	-	-	48.9	-	-	-	-
Cobra (Zhao et al., 2025)	3.5B	-	-	-	-	-	-	58.2	-	-	-	-
mmMamba-Hybrid (Liao et al., 2025)	2.7B	32.3	1371.1	63.7	-	-	-	55.1	-	-	-	-
MatVLM (Li et al., 2025b)	2.7B	37.3	1484.1	61.2	-	-	-	-	-	-	-	-
HYLIN	**3B**	**51.6**	**2121.0**	**79.8**	**56.0**	**80.8**	**82.6**	**79.1**	**92.4**	**74.8**	**807**	**63.3**
Larger LMMs (>4B)												
LLaVA-1.5 (Liu et al., 2024a)	7.4B	32.4	1510.7	64.3	-	-	-	-	-	-	-	-
MiniCPM-V2.6 (Yao et al., 2025)	8B	-	2348.4	78.0	57.5	82.1	82.4	80.1	90.8	-	852	65.0
LLaVA-SI (Li et al., 2025a)	7B	47.3	2109	81.7	60.9	81.6	78.8	78.5	86.9	65.3	622	65.5
Qwen2-VL (Wang et al., 2024c)	7B	54.1	2326.8	80.7	60.7	83.0	83.0	84.3	94.5	76.5	866	70.1
Qwen2.5-VL (Bai et al., 2025)	7B	58.6	2347.0	82.6	63.9	83.9	87.3	84.9	95.7	82.6	864	68.5
HYLIN	**7B**	**54.1**	**2243.0**	**83.4**	**63.9**	**82.1**	**85.7**	**82.7**	**94.2**	**80.3**	**865**	**67.9**

Table 3: **Video-understanding results.** Best linear/hybrid results are in **bold**.

Model	# Params	MVBench	LongVideoBench	MLVU	VideoMME	LVBench
Smaller Video LMMs (≤4B)						
Qwen2-VL (Wang et al., 2024c)	2B	63.2	-	-	55.6	-
Qwen2.5-VL (Bai et al., 2025)	3B	67.0	54.2	68.2	61.5	43.3
InternVL3 (Zhu et al., 2025)	2B	70.4	55.4	64.2	58.9	-
Smaller Linear Video LMMs (≤4B)						
AuroraLong (Xu et al., 2025b)	2B	53.2	-	52.7	-	-
HYLIN	**3B**	**67.5**	**52.7**	**67.6**	**60.6**	**43.1**
Larger Video LMMs (>4B)						
LLaVA-OneVision (Li et al., 2025a)	7B	56.7	56.3	64.7	58.2	-
LLaVA-Video (Zhang et al., 2024)	7B	58.6	58.2	70.8	63.3	-
Video-XL (Shu et al., 2025b)	7B	55.3	50.7	64.9	55.5	-
Qwen2-VL (Wang et al., 2024c)	7B	67.0	-	-	63.3	-
Qwen2.5-VL (Bai et al., 2025)	7B	69.6	56.0	70.2	65.1	45.3
Larger Linear Video LMMs (>4B)						
Vamba (Ren et al., 2025b)	10B	60.4	**55.9**	65.9	57.8	-
LongLLaVA (Wang et al., 2025c)	52B	64.6	53.5	-	53.8	-
HYLIN	**7B**	**70.1**	54.5	**69.6**	**64.1**	**45.1**

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

We apply HYLIN to Qwen2.5-VL-3B/7B (Bai et al., 2025), retaining their vision encoders and native dynamic-resolution support. The main VLMs are trained on 32 NVIDIA A800 GPUs with BF16 precision using 15M multimodal training examples. Runtime efficiency profiling uses 8 NVIDIA A800 GPUs (§3.3). Architectural ablations use direct SFT on a 3M-example image instruction subset of LLaVA-OneVision-1.5-Instruct, following the protocol described in Appendix F.1. Detailed per-stage training hyperparameters and the full evaluation benchmark list are provided in Appendix D.

3.2 MAIN RESULTS

Image Understanding. HYLIN maintains competitive performance across 11 image-understanding benchmarks with 25% full-attention layers (Table 2). After conversion and training, the 3B model exceeds the released Qwen2.5-VL-3B teacher on OCRBench (807 vs. 797) and MMBench-V1.1 (79.8 vs. 77.4), while scoring lower on MMMU and MME.

Video Understanding. With 25% full-attention layers, HYLIN-3B stays within one point of its teacher on MVBench, MLVU, and VideoMME (Table 3). HYLIN-7B outperforms Video-XL-7B on all four shared benchmarks and Vamba-10B on three, while trailing Vamba on LongVideoBench. In separate capacity tests, HYLIN supports 512+ frames on a single NVIDIA A800 80GB GPU, compared with approximately 256 for the teacher. Larger frame budgets also improve VideoMME accuracy (Appendix F.3).

Table 4: **Video-generation results** across attention-acceleration methods on HunyuanVideo-13B.

Method	Efficiency		VBench↑					
	Latency↓	Speedup↑	Quality Score	Semantic Score	Final Score	Imaging Quality	Aesthetic Quality	Subject Consist.
HunyuanVideo-13B (Text-to-Video, 720P, #Tokens 119K)								
FlashAttention-2 (Dao, 2024)	945s	1.00×	85.34	72.17	82.71	67.97	64.63	94.22
CLEAR (Liu et al., 2025)	2567s	0.37×	84.41	74.20	82.37	68.35	63.75	93.63
Swin (Liu et al., 2021)	762s	1.24×	80.91	71.35	79.00	62.22	51.01	91.74
STA (Zhang et al., 2025b)	527s	1.79×	84.63	73.83	82.47	68.39	64.18	94.75
SVG (Xi et al., 2025)	495s	1.91×	84.14	71.37	81.59	65.07	64.63	92.12
SVG2 (Yang et al., 2025a)	411s	2.30×	84.94	71.87	82.33	66.97	64.13	94.22
HYLIN	**486s**	**1.94×**	**85.17**	**73.62**	**82.86**	**68.24**	**64.41**	**94.38**

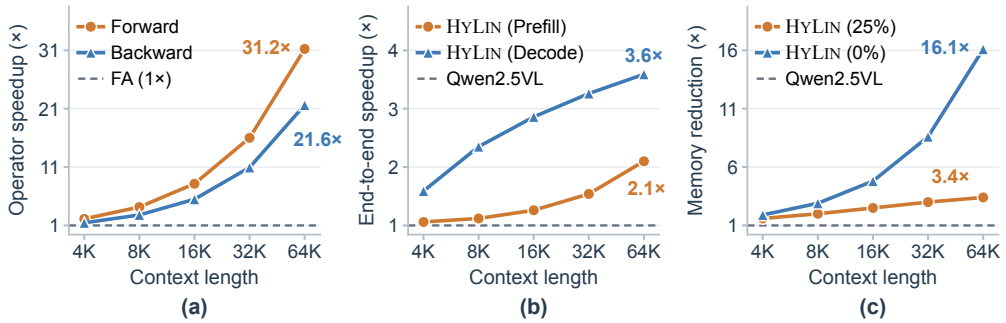


Figure 5: **Speed and memory comparison.** Profiling on 8 NVIDIA A800 GPUs relative to full-attention baselines: (a) HLWA forward/backward speedups; (b) HYLIN-7B end-to-end prefill/decode speedups; (c) GPU memory reductions with 25% and 0% retained FA layers.

Video Generation. On HunyuanVideo-13B at 720p with 119K tokens, our bidirectional DiT variant (§2.1) achieves the highest VBench Final score among the evaluated methods (82.86; Table 4). It reduces latency from 945 s to 486 s (1.94×) relative to FA2, with a similar Quality score and a higher Semantic score. Compared with the fastest baseline, SVG2, it gains 0.53 Final-score points at an additional cost of 75 s. Figure 9 shows qualitative examples.

3.3 RUNTIME EFFICIENCY ANALYSIS

Figure 5 profiles 4K–64K contexts on 8 NVIDIA A800 GPUs. At 64K, HLWA achieves 31.2× forward and 21.6× backward operator speedups over full attention. Against full-attention Qwen2.5-VL, HYLIN achieves 2.1× prefill and 3.6× decode speedups, with 3.4× lower peak memory at 25% retained FA (16.1× at 0%). At fixed head dimensions, HLWA uses a fixed-size visual state and a full causal text KV cache, requiring $\mathcal{O}(N_t)$ cache memory and $\mathcal{O}(N_t^2)$ text-prefill cost. Retained FA layers still require $\mathcal{O}(N)$ cache memory and $\mathcal{O}(N^2)$ prefill, where $N = N_v + N_t$. The asymptotic complexity is unchanged, and efficiency gains depend on the visual/text token composition.

4 CONCLUSION

In this work, we present HYLIN, a hybrid attention framework centered on computation allocation for multimodal understanding and generation. HLWA combines global visual linear attention with causal attention over the full text prefix within converted VLM layers, while operator error and task sensitivity guide full-attention retention for cross-modal interaction. With 75% of VLM layers converted, HYLIN-3B closely tracks its teacher on the reported image and video understanding benchmarks. At 64K tokens, HYLIN-7B delivers 2.1× prefill and 3.6× decode speedups with a 3.4× reduction in peak GPU memory over its full-attention teacher. A bidirectional variant further accelerates HunyuanVideo by 1.94× with competitive generation quality. By coordinating global compression, exact token-level retrieval, and selective full attention, HYLIN offers a foundation for more capable and scalable multimodal architectures.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, manuscript organization, and the presentation and checking of mathematical arguments. The authors reviewed and independently verified all AI-assisted text and derivations. Generative AI was not used to fabricate experimental data or reported measurements. We take responsibility for the final content, including all claims, proofs, results, and artifacts.

ETHICS STATEMENT

This work studies efficient attention mechanisms for vision-language understanding and video generation, and relies only on publicly available benchmarks and datasets used in accordance with their licenses. It does not involve human subjects, private data, or the release of new data. As with any efficient generative or multimodal model, care should be taken to avoid misuse (e.g., generation of misleading media); our contributions target computational efficiency and do not introduce new capabilities beyond those of the underlying teacher models.

REPRODUCIBILITY STATEMENT

The VLM attention operator, layer-selection criterion, and four-stage training procedure are described in section 2, with proofs of the error statements provided in Appendix C. Training hyperparameters and evaluation benchmarks are summarized in Appendix D. Appendix F.1 reports measured comparisons of layer-selection strategies and sensitivity to λ under a matched full-attention budget, initialization, and direct-SFT protocol. The bidirectional Video-DiT formulation and adaptation setup are outlined in §2.1 and Appendix D.2, respectively.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, Haozhe Liu, Hongwei Yi, Hao Zhang, Muyang Li, Yukang Chen, Han Cai, Sanja Fidler, Ping Luo, Song Han, and Enze Xie. SANA-Video: Efficient Video Generation with Block Linear Diffusion Transformer. In *International Conference on Learning Representations*, 2026.

Lin Chen, Zehui Chen, Xiaoyi Dong, Haodong Duan, Jinsong Li, Dahua Lin, Yu Qiao, Jiaqi Wang, Yuhang Zang, Pan Zhang, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, volume 2024, pp. 35549–35562, 2024.

Tri Dao and Albert Gu. Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 10041–10071, 2024.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models. *arXiv preprint arXiv:2402.19427*, 2024.

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van Keirsbilck, Min-Hung Chen, Yoshi Suhara, Yingyan Lin, Jan Kautz,

and Pavlo Molchanov. Hymba: A hybrid-head architecture for small language models. In *International Conference on Learning Representations*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025a. doi: 10.52202/085713-4899.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-Modal LLMs in Video Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025b.

Mohsen Ghafoorian, Denis Korzhenkov, and Amirhossein Habibian. Attention Surgery: An Efficient Recipe to Linearize Your Video Diffusion Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32915–32925, 2026.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, Anna Golubeva, Vasudev Shyam, James Whittington, Jonathan Pilault, and Beren Millidge. The Zamba2 Suite: Technical Report. *arXiv preprint arXiv:2411.15242*, 2024.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. MInference 1.0: Accelerating Pre-filling for Long-Context LLMs via Dynamic Sparse Attention. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1663.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5156–5165, 2020.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram Is Worth a Dozen Images. In *European Conference on Computer Vision (ECCV)*, pp. 235–251, 2016.

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*, 2024.

Disen Lan, Weigao Sun, Jiaxi Hu, Jusen Du, and Yu Cheng. Liger: Linearizing Large Language Models to Gated Recurrent Structures. In *International Conference on Machine Learning (ICML)*, pp. 32452–32466, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*, 2025a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206. IEEE, 2024.

Yingyue Li, Bencheng Liao, Wenyu Liu, and Xinggang Wang. MatVLM: Hybrid Mamba-Transformer for Efficient Vision-Language Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20878–20888. IEEE, 2025b.

Bencheng Liao, Hongyuan Tao, Qian Zhang, Tianheng Cheng, Yingyue Li, Haoran Yin, Wenyu Liu, and Xinggang Wang. Multimodal mamba: Decoder-only multimodal state space model via quadratic to linear distillation. *arXiv preprint arXiv:2502.13145*, 2025.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A Hybrid Transformer-Mamba Language Model. *arXiv preprint arXiv:2403.19887*, 2024.

Anmin Liu, Ruixuan Yang, Huiqiang Jiang, Bin Lin, Minmin Sun, Yong Li, Chen Zhang, and Tao Xie. VecAttention: Vector-wise Sparse Attention for Accelerating Long Context Inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41299–41310, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1516.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Songhua Liu, Zhenxiong Tan, and Xinchao Wang. CLEAR: Conv-Like Linearization Revs Pre-Trained Diffusion Transformers Up. In *Advances in Neural Information Processing Systems*, volume 38, pp. 154623–154657, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-Around Player? In *Computer Vision – ECCV 2024*, pp. 216–233, 2024b.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models. *Science China Information Sciences*, 67(12):220102, 2024c. doi: 10.1007/s11432-024-4235-6.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. DeepSeek-VL: Towards Real-World Vision-Language Understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-InternVL: Pushing the Boundaries of Monolithic Multimodal Large Language Models with Endogenous Visual Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24960–24971. IEEE, 2025.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A Dataset for VQA on Document Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1697–1706, 2022.

Jean Mercat, Igor Vasiljevic, Sedrick Keh, Kushal Arora, Achal Dave, Adrien Gaidon, and Thomas Kollar. Linearizing Large Language Models. *arXiv preprint arXiv:2405.06640*, 2024.

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, Enwei Jiao, Gengxin Li, Guojun Zhang, Haohai Sun, Houze Dong, Jiadai Zhu, Jiaqi Zhuang, Jiayuan Song, Jin Zhu, Jingtao Han, Jingyang Li, Junbin Xie, Junhao Xu, Junjie Yan, Kaishun Zhang, Kecheng Xiao, Kexi Kang, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Zheng, Linbo Chai, Long Xing, Meizhi Ju, Mingyuan Chi, Mozhi Zhang, Peikai Huang, Pengcheng Niu, Pengfei Li, Pengyu Zhao, Qi Yang, Qidi Xu, Qiexiang Wang, Qin Wang, Qiuhui Li, Ruitao Leng, Shengmin Shi, Shuqi Yu, Sichen Li, Songquan Zhu, Tao Huang, Tianrun Liang, Weigao Sun, Weixuan Sun, Weiyu Cheng, Wenkai Li, Xiangjun Song, Xiao Su, Xiaodong Han, Xinjie Zhang, Xinzhu Hou, Xu Min, Xun Zou, Xuyang Shen, Yan Gong, Yingjie Zhu, Yipeng Zhou, Yiran Zhong, Yongyi Hu, Yuanxiang Fan, Yue Yu, Yufeng Yang, Yuhao Li, Yunan Huang, Yunji Li, Yunpeng Huang, Yunzhi Xu, Yuxin Mao, Zehan Li, Zekang Li, Zewei Tao, Zewen Ying, Zhaoyang Cong, Zhen Qin, Zhenhua Fan, Zhihang Yu, Zhuo Jiang, and Zijia Wu. MiniMax-01: Scaling Foundation Models with Lightning Attention. *arXiv preprint arXiv:2501.08313*, 2025.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. VL-Mamba: Exploring State Space Models for Multimodal Learning. *arXiv preprint arXiv:2403.13600*, 2024.

Zhen Qin, Songlin Yang, and Yiran Zhong. Hierarchically Gated Recurrent Neural Network for Sequence Modeling. In *Advances in Neural Information Processing Systems*, volume 36, pp. 33202–33221, 2023.

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling. In *International Conference on Learning Representations*, 2025a.

Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21197–21208. IEEE, 2025b.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Bala Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 54582–54599, 2025.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-XL: Extra-Long Vision Language Model for Hour-Scale Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26160–26169. IEEE, 2025a.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-XL: Extra-Long Vision Language Model for Hour-Scale Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26160–26169. IEEE, 2025b.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8317–8326, 2019.

Hongyuan Tao, Bencheng Liao, Shaoyu Chen, Haoran Yin, Qian Zhang, Wenyu Liu, and Xinggang Wang. InfiniteVL: Synergizing Linear and Sparse Attention for Highly-Efficient, Unlimited-Input Vision-Language Models. *arXiv preprint arXiv:2512.08829*, 2025.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, Wentao Li, Enzhe Lu, Weizhou Liu, Yanru Chen, Weixin Xu, Longhui Yu, Yejie Wang, Yu Fan, Longguang Zhong, Enming Yuan, Dehao Zhang, Yizhi Zhang, T. Y. Liu, Haiming Wang, Shengjun Fang, Weiran He, Shaowei Liu, Yiwei Li, Jianlin Su, Jiezhong Qiu, Bo Pang, Junjie Yan, Zhejun Jiang, Weixiao Huang, Bohong Yin, Jiacheng You, Chu Wei, Zhengtao Wang, Chao Hong, Yutian Chen, Guanduo Chen, Yucheng Wang, Huabin Zheng, Feng Wang, Yibo Liu, Mengnan Dong, Zheng Zhang, Siyuan Pan, Wenhao Wu, Yuhao Wu, Longyu Guan, Jiawen Tao, Guohong Fu, Xinran Xu, Yuzhi Wang, Guokun Lai, Yuxin Wu, Xinyu Zhou, Zhilin Yang, and Yulun Du. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.

Dezhan Tu, Danylo Vashchilenko, Yuzhe Lu, and Panpan Xu. VL-Cache: Sparsity and Modality-Aware KV Cache Compression for Vision-Language Model Inference Acceleration. In *International Conference on Learning Representations*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Q-VLM: Post-training Quantization for Large Vision-Language Models. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-3638.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024c.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. LVBench: An Extreme Long Video Understanding Benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou,

Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Xidong Wang, Dingjie Song, Shunian Chen, Junying Chen, Zhenyang Cai, Chen Zhang, Lichao Sun, and Benyou Wang. LongLLaVA: Scaling multi-modal LLMs to 1000 images efficiently via a hybrid architecture. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21419–21436, Suzhou, China, 2025c. Association for Computational Linguistics.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. doi: 10.52202/079017-0907.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, Jianfei Chen, Ion Stoica, Kurt Keutzer, and Song Han. Sparse VideoGen: Accelerating Video Diffusion Transformers with Spatial-Temporal Sparsity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68208–68224. PMLR, 2025.

Boshen Xu, Zihan Xiao, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Jian Luan, and Qin Jin. TimeViper: A Hybrid Mamba-Transformer Vision-Language Model for Efficient Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21784–21794, 2026.

Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. XAttention: Block Sparse Attention with Antidiagonal Scoring. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69819–69831. PMLR, 2025a.

Weili Xu, Enxin Song, Wenhao Chai, Xuexiang Wen, Tian Ye, and Gaoang Wang. Bringing RNNs Back to Efficient Open-Ended Video Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23453–23465, 2025b.

Yufei Xue, Yushi Huang, Jiawei Shao, Lunjie Zhu, Chi Zhang, Xuelong Li, and Jun Zhang. VLMQ: Token Saliency-Driven Post-Training Quantization for Vision-language Models. *arXiv preprint arXiv:2508.03351*, 2025.

Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, Jianfei Chen, Song Han, Kurt Keutzer, and Ion Stoica. Sparse VideoGen2: Accelerate Video Generation with Sparse Attention via Semantic-Aware Permutation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 96965–96991, 2025a.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated Linear Attention Transformers with Hardware-Efficient Training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56501–56523. PMLR, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule. In *International Conference on Learning Representations*, 2025b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, 2025c.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*, 2024.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Chi Chen, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Ronghua Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Efficient GPT-4V level multimodal large language model for deployment on edge devices. *Nature Communications*, 2025. doi: 10.1038/s41467-025-61040-5.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567. IEEE, 2024.

Michael Zhang, Simran Arora, Rahul Chalamala, Benjamin Spector, Alan Wu, Krithik Ramesh, Aaryan Singhal, and Christopher Ré. LoLCaTs: On Low-Rank Linearizing of Large Language Models. In *International Conference on Learning Representations*, 2025a.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast Video Generation with Sliding Tile Attention. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 74714–74731, 2025b.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token. In *International Conference on Learning Representations*, 2025c.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video Instruction Tuning With Synthetic Data. *arXiv preprint arXiv:2410.02713*, 2024.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10421–10429, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking Multi-task Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, June 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025.

Jingwei Zuo, Maksim Velikanov, Ilyas Chahed, Younes Belkada, Dhia Eddine Rhayem, Guillaume Kunsch, Hakim Hacid, Hamza Yous, Brahim Farhat, Ibrahim Khadraoui, Mugariya Farooq, Giulia Campesan, Ruxandra Cojocaru, Yasser Djilali, Shi Hu, Iheb Chaabane, Puneesh Khanna, Mohamed El Amine Seddik, Ngoc Dung Huynh, Phuc Le Khac, Leen AlQadi, Billel Mokeddem, Mohamed Chami, Abdalgader Abubaker, Mikhail Lubinets, Kacper Piskorski, and Slim Frikha. Falcon-h1: A family of hybrid-head language models redefining efficiency and performance. *arXiv preprint arXiv:2507.22448*, 2025.

Appendix

A Related Work

Efficient attention for long-context multimodal models. Numerous efforts have tried to break the $\mathcal{O}(N^2)$ cost of softmax attention. One line of work replaces softmax with a recurrent state at $\mathcal{O}(N)$ cost: linear attention (Katharopoulos et al., 2020) and gated variants such as GLA (Yang et al., 2024), Mamba-2 (Dao & Gu, 2024), HGRN (Qin et al., 2023) and Gated DeltaNet (Yang et al., 2025b) are now standard backbones for efficient sequence modelling, yet they consistently lose token-level retrieval relative to softmax (De et al., 2024; Ren et al., 2025a). A second line keeps softmax but reduces the input it sees, through token pruning and merging (Shen et al., 2025; Zhang et al., 2025c; Shu et al., 2025a), KV-cache compression and quantisation (Tu et al., 2025; Wang et al., 2024a; Xue et al., 2025), or sparse-attention selectors (Jiang et al., 2024; Xu et al., 2025a; Liu et al., 2026; Xi et al., 2025; Ghafoorian et al., 2026; Xu et al., 2026). A third line combines the two into hybrid linear-softmax architectures, mixing the operators either across layers (a 3:1 linear-to-softmax ratio is now standard (Lieber et al., 2024; Glorioso et al., 2024; MiniMax et al., 2025; Team et al., 2025; Li et al., 2025b; Ren et al., 2025b; Wang et al., 2025c)) or within a single layer (Ren et al., 2025a; De et al., 2024; Dong et al., 2025; Zuo et al., 2025; Zhang et al., 2025a; Lan et al., 2025). InfiniteVL (Tao et al., 2025), for example, interleaves Gated DeltaNet and sliding-window attention layers, applying the same token-mixing operator to visual and text tokens within each layer. Intra-layer hybrids typically combine both pathways over the full token sequence. In HyLin, each converted layer routes visual queries to Gated DeltaNet and text queries to causal softmax attention over the full text prefix, without a fixed window cap. HyLin further selects the retained FA layers jointly by calibrated operator error and downstream task sensitivity, providing direct cross-modal interaction at those layers.

Vision-Language Models and video Diffusion Transformers. Two multimodal paradigms motivate this work, and both are bottlenecked by full softmax attention. *Vision-Language Models* concatenate visual embeddings from a pre-trained vision encoder with text embeddings and process the joint sequence through a decoder-only LLM backbone, a recipe established by LLaVA (Liu et al., 2023) and scaled by InternVL (Wang et al., 2025b; Zhu et al., 2025), Qwen-VL (Wang et al., 2024b; Bai et al., 2025), DeepSeek-VL (Lu et al., 2024) and MiniCPM-V (Yao et al., 2024); native dynamic-resolution support and 2D Multimodal RoPE in Qwen2.5-VL (Bai et al., 2025) are essential for fine-grained document and chart parsing, and we inherit both. *Video Diffusion Transformers (DiTs)* (Peebles & Xie, 2023) model video generation as iterative denoising over a 3D latent grid of patch tokens, using bidirectional self-attention over the entire spatial-temporal sequence at every denoising step. Systems such as CogVideoX (Yang et al., 2025c), HunyuanVideo (Kong et al., 2024) and Wan2.1 (Wan et al., 2025) therefore incur quadratic attention repeatedly during sampling. SANA-Video (Chen et al., 2026) demonstrates that bidirectional ReLU Linear Attention with 3D RoPE can scale global video mixing efficiently, but a pure linear operator can still lose fine-grained local interactions. This motivates our complementary local branch and the shared principle of combining global compression with exact retrieval.

Knowledge distillation for linearisation. Transferring a full-attention teacher into an efficient student is the standard recipe for hybrid models. LoLCATs (Zhang et al., 2025a) aligns student and teacher outputs with MSE on top of LoRA adapters; SUPRA (Mercat et al., 2024) performs full-parameter linearisation; MiniLM (Wang et al., 2020) distils the structural information encoded in attention maps. HyLin adapts distillation to modality-aware attention replacement with a four-stage pipeline that combines layer-wise output alignment, importance-based attention layer selection, logit-level distillation, and text-supervised fine-tuning, enabling robust transfer in the multimodal setting.

B Preliminaries

Architecture Overview In Vision-Language Models (VLMs), the input is a multimodal sequence $\mathbf{X}^{(0)} \in \mathbb{R}^{N \times d}$ whose row indices are partitioned into visual tokens $\mathcal{V}$ and text-side tokens $\mathcal{T}$, with $N = N_v + N_t$. The common contiguous layout $\mathbf{X}^{(0)} = [\mathbf{E}_{\text{vision}}; \mathbf{E}_{\text{text}}]$ is a special case; the

formulation also permits interleaved layouts. Both full-attention and hybrid VLMs consist of a stack of L layers. The forward pass of the l-th layer can be formalized as:

$$\begin{aligned}\mathbf{H}^{(l)} &= \text{Mixer}^{(l)}\left(\mathbf{X}^{(l-1)}\right) + \mathbf{X}^{(l-1)},\\ \mathbf{X}^{(l)} &= \text{FFN}^{(l)}\left(\mathbf{H}^{(l)}\right) + \mathbf{H}^{(l)},\end{aligned} \tag{19}$$

where $\mathbf{X}^{(l)} \in \mathbb{R}^{N\times d}$ denotes the output embeddings of the l-th layer, and FFN is the feed-forward network. In a hybrid VLM architecture, the set of retained full-attention layers is specified by $\mathcal{I}_{\text{FA}} \subset \{1,\dots,L\}$. The sequence mixers are defined according to the layer index:

$$\text{Mixer}^{(l)} = \begin{cases}\text{Softmax-Attn}^{(l)} & \text{if } l \in \mathcal{I}_{\text{FA}},\\ \text{HLWA}^{(l)} & \text{otherwise.}\end{cases} \tag{20}$$

Standard Attention in VLMs In standard Transformer-based VLMs, the mixer layer uses softmax attention to capture cross-modal interactions. Omitting the multi-head mechanism for simplicity, it can be written as:

$$\begin{aligned}&\mathbf{Q} = \mathbf{X}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_k, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_v,\\ &\mathbf{O} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}\mathbf{W}_o,\end{aligned} \tag{21}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d\times d_k}$ and $\mathbf{W}_o \in \mathbb{R}^{d_k\times d}$ are learnable projection matrices. While highly expressive, the $\mathcal{O}(N^2)$ computational complexity of softmax attention becomes a prohibitive bottleneck when scaling to high-resolution images or long-context videos.

Linear Attention and Recurrent Formulation Linear attention simplifies the traditional attention mechanism by replacing the exponential term $\exp(\mathbf{q}_t\mathbf{k}_i^\top)$ with a kernel function $k(\mathbf{x},\mathbf{y})$ (Katharopoulos et al., 2020; Qin et al., 2023), typically defined via an associated feature map ϕ such that $k(\mathbf{x},\mathbf{y}) = \langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle$.

To mitigate the quadratic complexity, modern Linear Attention and RNN variants formulate the sequence mixing as a recurrent process. For a given time step t, the computation can be written as:

$$\mathbf{q}_t = \phi(\mathbf{x}_t\mathbf{W}_q), \quad \mathbf{k}_t = \phi(\mathbf{x}_t\mathbf{W}_k), \quad \mathbf{v}_t = \mathbf{x}_t\mathbf{W}_v, \tag{22}$$

$$\mathbf{S}_t = \alpha_t\mathbf{S}_{t-1} + \beta_t\mathbf{v}_t\mathbf{k}_t^\top \in \mathbb{R}^{d_v\times d_k}, \tag{23}$$

$$\mathbf{o}_t = \mathbf{S}_t\mathbf{q}_t \in \mathbb{R}^{d_v}, \tag{24}$$

where $\phi(\cdot)$ is the kernel feature map, and $\mathbf{S}_t$ is the matrix-valued *recurrent state*. Eq. (23) and (24) are referred to as the *update rule* and the *query rule*, respectively.

Modern variants parameterize α_t and β_t with different scalar-, channel-, or state-dependent gates. These gates control how past state is forgotten and how new key–value information is written; the exact parameterization follows each original implementation.

Representative updates are summarized in Appendix Table 5. This formulation achieves $\mathcal{O}(Nd_kd_v)$ complexity, making it efficient for long multimodal sequences. We compare these linear mixers under the same HLWA routing and training pipeline.

C Additional HLWA Details and Proofs

We follow the convention of §2.1: all row-level symbols below are evaluated at a fixed layer and head, whose indices are suppressed.

C.1 Visual Linear-Attention State Updates

Table 5 summarizes the recurrent state updates and readout rules for the four linear-attention variants considered in this work. All models maintain a fixed-size hidden state $\mathbf{S}_t$ that is updated at each visual token and read out via a query projection. HGRN (Qin et al., 2023) uses a scalar gate α_t with an element-wise EMA over values, yielding a vector-valued state $\mathbf{S}_t \in \mathbb{R}^{d_v}$ that is read out by an

element-wise query modulation. GLA (Yang et al., 2024) and Mamba-2 (Dao & Gu, 2024) generalize to matrix-valued states $\mathbf{S}_t \in \mathbb{R}^{d_v \times d_k}$ by accumulating outer products $\mathbf{v}_t \mathbf{k}_t^\top$, differing in whether the decay is element-wise (GLA) or scalar (Mamba-2). Gated DeltaNet (Yang et al., 2025b) further introduces a delta rule that erases the old key direction before writing ($\mathbf{I} - \beta_t \mathbf{k}_t \mathbf{k}_t^\top$), improving in-context associative recall. We adopt Gated DeltaNet as the default visual LA operator in HYLIN due to its favorable trade-off between state expressiveness and convergence speed observed during Stage-1 distillation.

Table 5: Representative state updates and readouts for visual linear-attention variants.

Model	**Representative state update**	**Readout rule**
HGRN (Qin et al., 2023)	$\mathbf{S}_t = \alpha_t \mathbf{S}_{t-1} + (1-\alpha_t)\mathbf{v}_t \in \mathbb{R}^{d_v}$	$\mathbf{o}_t = \mathbf{S}_t \odot \mathbf{q}_t \in \mathbb{R}^{d_v}$
GLA (Yang et al., 2024)	$\mathbf{S}_t = \mathbf{S}_{t-1} \odot (\mathbf{1}\alpha_t^\top) + \mathbf{v}_t \mathbf{k}_t^\top \in \mathbb{R}^{d_v \times d_k}$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t \in \mathbb{R}^{d_v}$
Mamba-2 (Dao & Gu, 2024)	$\mathbf{S}_t = \alpha_t \mathbf{S}_{t-1} + \mathbf{v}_t \mathbf{k}_t^\top \in \mathbb{R}^{d_v \times d_k}$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t \in \mathbb{R}^{d_v}$
Gated DeltaNet (Yang et al., 2025b)	$\mathbf{S}_t = \alpha_t \mathbf{S}_{t-1}(\mathbf{I} - \beta_t \mathbf{k}_t \mathbf{k}_t^\top) + \beta_t \mathbf{v}_t \mathbf{k}_t^\top \in \mathbb{R}^{d_v \times d_k}$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t \in \mathbb{R}^{d_v}$

C.2 VISUAL COMPRESSION CERTIFICATES

Centered LogSumExp and Softmax expansion. Let $\delta_j = z_j - \bar{z}$, $\rho = \|\boldsymbol{\delta}\|_\infty$, and $\mu_m = n_i^{-1} \sum_j \delta_j^m$. Since $\mu_1 = 0$, expanding the exponential and then $\log(1+x)$ gives

$$\begin{aligned} J(\mathbf{z}) &= \log\left(\frac{1}{n_i}\sum_j e^{\delta_j}\right) \\ &= \log\Big(1 + \frac{\mu_2}{2} + \frac{\mu_3}{6} + \mathcal{O}(\rho^4)\Big) \\ &= \frac{\mu_2}{2} + \frac{\mu_3}{6} + \mathcal{O}(\rho^4) = \frac{1}{2}\operatorname{Var}_{\mathbf{u}}(\mathbf{z}) + \mathcal{O}(\rho^3), \end{aligned} \tag{25}$$

which proves Eq. (3). Using $\mathrm{Softmax} = \nabla \mathrm{LSE}$, or directly expanding $e^{\delta_j} / \sum_r e^{\delta_r}$, also gives

$$p_{ij} = \frac{1}{n_i}\left[1 + \delta_j + \frac{1}{2}\left(\delta_j^2 - \mu_2\right)\right] + \mathcal{O}\left(\frac{\rho^3}{n_i}\right). \tag{26}$$

Each displayed truncation sums to one, although a finite truncation need not remain componentwise nonnegative.

For the row-wise state form, let $n_i = |\mathcal{A}_i|$, $\bar{z}_i = n_i^{-1}\sum_j z_{ij}$, $\delta_{ij} = z_{ij} - \bar{z}_i$, $\bar{\mathbf{v}}_i = n_i^{-1}\sum_j \mathbf{v}_j$, and $\bar{\mathbf{k}}_i = n_i^{-1}\sum_j \mathbf{k}_j$. Define

$$\mathbf{C}_i = \frac{1}{n_i\sqrt{d_k}} \sum_{j \in \mathcal{A}_i} (\mathbf{v}_j - \bar{\mathbf{v}}_i)(\mathbf{k}_j - \bar{\mathbf{k}}_i)^\top. \tag{27}$$

Because $\delta_{ij} = \mathbf{q}_i^\top(\mathbf{k}_j - \bar{\mathbf{k}}_i)/\sqrt{d_k}$ and $\sum_j \delta_{ij} = 0$, the first-order output in Proposition 1 has the fixed-size state form

$$\mathbf{o}_i^{(1)} = \bar{\mathbf{v}}_i + \mathbf{C}_i \mathbf{q}_i. \tag{28}$$

Proof of Proposition 1. For a non-asymptotic first-order remainder, define the path

$$p_j(a) = \frac{e^{a\delta_j}}{\sum_r e^{a\delta_r}}, \qquad m(a) = \sum_j p_j(a)\delta_j, \qquad \sigma^2(a) = \sum_j p_j(a)(\delta_j - m(a))^2. \tag{29}$$

Direct differentiation yields

$$p_j'(a) = p_j(a)(\delta_j - m(a)), \qquad p_j''(a) = p_j(a)\left[(\delta_j - m(a))^2 - \sigma^2(a)\right]. \tag{30}$$

Because all $\delta_j \in [-\rho_i, \rho_i]$, Popoviciu's inequality gives $\sigma^2(a) \le \rho_i^2$, and therefore

$$\|\mathbf{p}''(a)\|_1 \le \sum_j p_j(a)\left[(\delta_j - m(a))^2 + \sigma^2(a)\right] = 2\sigma^2(a) \le 2\rho_i^2. \tag{31}$$

The integral form of Taylor's theorem now gives

$$p_{ij} = \frac{1+\delta_{ij}}{n_i} + r_{ij}, \qquad \mathbf{r}_i = \int_0^1 (1-a)\mathbf{p}''(a)\,da, \qquad \|\mathbf{r}_i\|_1 \le \rho_i^2. \tag{32}$$

Both $\mathbf{p}_i$ and its first-order truncation sum to one, so $\sum_j r_{ij} = 0$. Moreover,

$$\mathbf{C}_i\mathbf{q}_i = \frac{1}{n_i}\sum_j \delta_{ij}(\mathbf{v}_j - \bar{\mathbf{v}}_i) = \frac{1}{n_i}\sum_j \delta_{ij}\mathbf{v}_j. \tag{33}$$

Consequently,

$$\mathbf{o}_i^{\mathrm{FA}} - \bar{\mathbf{v}}_i - \mathbf{C}_i\mathbf{q}_i = \sum_j r_{ij}\mathbf{v}_j. \tag{34}$$

The positive and negative parts of $\mathbf{r}_i$ have the same mass $\gamma = \|\mathbf{r}_i\|_1/2$. After normalization, they define two convex combinations of the value vectors, whose distance is at most D_i. Hence

$$\left\|\sum_j r_{ij}\mathbf{v}_j\right\|_2 \le \frac{D_i}{2}\|\mathbf{r}_i\|_1 \le \frac{D_i}{2}\rho_i^2, \tag{35}$$

which proves Eq. (4). □

For completeness, if $\psi(a) = \log(n_i^{-1}\sum_j e^{a\delta_j})$, then the entropy deficit satisfies $\Delta = \psi'(1) - \psi(1)$ and therefore

$$\Delta = \frac{1}{2}\operatorname{Var}_{\mathbf{u}}(\mathbf{z}) + \mathcal{O}(\rho^3). \tag{36}$$

These expansions are used only as calibration diagnostics; exact LogSumExp and entropy statistics are used in all reported measurements.

Entropy-deficit certificate.

Proof of Proposition 2. Let $\mathbf{u}_i$ be uniform on $\mathcal{A}_i$. The entropy deficit is exactly

$$D_{\mathrm{KL}}(\mathbf{p}_i\|\mathbf{u}_i) = \sum_j p_{ij}\log(n_i p_{ij}) = \log n_i - H(\mathbf{p}_i) = \Delta_i. \tag{37}$$

For $a_j = p_{ij} - 1/n_i$, we have $\sum_j a_j = 0$. Let $\gamma = \sum_{j:a_j>0} a_j = \frac{1}{2}\sum_j |a_j| = \mathrm{TV}(\mathbf{p}_i, \mathbf{u}_i)$. When $\gamma > 0$, the normalized positive and negative parts of $\{a_j\}$ define two convex combinations $\mathbf{v}^+$ and $\mathbf{v}^-$ of the value vectors, and hence

$$\left\|\mathbf{o}_i^{\mathrm{FA}} - \bar{\mathbf{v}}_i\right\|_2 = \gamma\|\mathbf{v}^+ - \mathbf{v}^-\|_2 \le D_i\,\mathrm{TV}(\mathbf{p}_i, \mathbf{u}_i). \tag{38}$$

Pinsker's inequality gives $\mathrm{TV}(\mathbf{p}_i, \mathbf{u}_i) \le \sqrt{D_{\mathrm{KL}}(\mathbf{p}_i\|\mathbf{u}_i)/2}$. Combining this with $\mathrm{TV}(\mathbf{p}_i, \mathbf{u}_i) \le 1$ gives the minimum in Eq. (5). The case $\gamma = 0$ is immediate. □

C.3 Visual Conditioning and Replacement Error

Fix a layer, head, teacher hidden state, and attention mask. For a visual query $i \in \mathcal{V}$, partition its legal keys as

$$\mathcal{A}_i^V = \mathcal{A}_i \cap \mathcal{V}, \qquad \mathcal{A}_i^T = \mathcal{A}_i \cap \mathcal{T}, \qquad \tau_i^{v\to t} = \sum_{j\in\mathcal{A}_i^T} p_{ij}. \tag{39}$$

Assume finite logits and $\mathcal{A}_i^V \neq \varnothing$, so $0 \le \tau_i^{v\to t} < 1$. Conditioning the teacher distribution gives

$$p_{ij}^V = \frac{p_{ij}}{1-\tau_i^{v\to t}} = \frac{e^{z_{ij}}}{\sum_{r\in\mathcal{A}_i^V} e^{z_{ir}}}, \qquad \boldsymbol{\mu}_i^V = \sum_{j\in\mathcal{A}_i^V} p_{ij}^V\mathbf{v}_j. \tag{40}$$

For $\tau_i^{v\to t} > 0$, define $p_{ij}^T = p_{ij}/\tau_i^{v\to t}$ on $\mathcal{A}_i^T$ and $\boldsymbol{\mu}_i^T = \sum_{j\in\mathcal{A}_i^T} p_{ij}^T\mathbf{v}_j$. When $\tau_i^{v\to t} = 0$, every omission term multiplied by this mass is defined as zero, without requiring a mean over an empty text set. A self-inclusive causal mask ensures that $\mathcal{A}_i^V$ is nonempty for a visual query. The conditional statements below do not apply to a mask with $\mathcal{A}_i^V = \varnothing$.

Omission and compression. Partitioning the teacher output yields $\mathbf{o}_i^{\mathrm{FA}} = (1-\tau_i^{v\to t})\boldsymbol{\mu}_i^V + \tau_i^{v\to t}\boldsymbol{\mu}_i^T$. Adding and subtracting $\boldsymbol{\mu}_i^V$ therefore gives the exact identity

$$\mathbf{o}_i^{\mathrm{FA}} - \mathbf{o}_i^{\mathrm{LA}} = \underbrace{\tau_i^{v\to t}(\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_i^V)}_{\mathbf{a}_i} + \underbrace{(\boldsymbol{\mu}_i^V - \mathbf{o}_i^{\mathrm{LA}})}_{\mathbf{b}_i}, \tag{41}$$

where $\mathbf{o}_i^{\mathrm{LA}}$ is the actual visual-only mixer output on the same teacher input. The two vectors represent omitted direct text access and approximation of the conditional visual output, respectively. Writing $\eta_i^V = \|\mathbf{b}_i\|_2$, the triangle inequality proves Eq. (6) and gives

$$\epsilon_i^{\mathrm{LA}} \le \tau_i^{v\to t}\|\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_i^V\|_2 + \eta_i^V \le \tau_i^{v\to t} D_i + \eta_i^V, \tag{42}$$

where D_i is the diameter over all legal values. Both conditional means are convex combinations of those values when their masses are positive. The squared error retains a cross term:

$$(\epsilon_i^{\mathrm{LA}})^2 = \|\mathbf{a}_i\|_2^2 + \|\mathbf{b}_i\|_2^2 + 2\langle \mathbf{a}_i, \mathbf{b}_i\rangle. \tag{43}$$

No orthogonality is assumed, so the two squared contributions cannot in general be added independently.

Compression diagnostics on visual keys. To apply Propositions 1 and 2 to $\boldsymbol{\mu}_i^V$, recompute the statistics inside $\mathcal{A}_i^V$:

$$n_i^V = |\mathcal{A}_i^V|, \qquad \bar{z}_i^V = \frac{1}{n_i^V}\sum_{j\in\mathcal{A}_i^V} z_{ij}, \qquad \delta_{ij}^V = z_{ij} - \bar{z}_i^V, \tag{44}$$

$$\rho_i^V = \max_{j\in\mathcal{A}_i^V} |\delta_{ij}^V|, \qquad \Delta_i^V = \log n_i^V - H(\mathbf{p}_i^V), \tag{45}$$

$$D_i^V = \mathrm{diam}\{\mathbf{v}_j : j \in \mathcal{A}_i^V\}. \tag{46}$$

Let $\bar{\mathbf{v}}_i^V = (n_i^V)^{-1}\sum_{j\in\mathcal{A}_i^V}\mathbf{v}_j$ and $\mathbf{o}_i^{(1,V)} = (n_i^V)^{-1}\sum_{j\in\mathcal{A}_i^V}(1+\delta_{ij}^V)\mathbf{v}_j$. Since $\mathbf{p}_i^V$ is precisely softmax on the restricted logits, the same proofs give

$$\|\boldsymbol{\mu}_i^V - \mathbf{o}_i^{(1,V)}\|_2 \le \frac{D_i^V}{2}(\rho_i^V)^2, \tag{47}$$

$$\|\boldsymbol{\mu}_i^V - \bar{\mathbf{v}}_i^V\|_2 \le D_i^V \min\left\{1, \sqrt{\frac{\Delta_i^V}{2}}\right\}. \tag{48}$$

These certificates concern the specified linear surrogate and uniform value aggregation. They do not directly bound η_i^V for a learned Gated DeltaNet, and the all-key statistics ρ_i and Δ_i cannot simply replace the conditional statistics above. Nor does high entropy control the omission term: with one visual value 0, one text value 1, and equal logits, $\rho_i = \Delta_i = 0$ but $\mathbf{o}_i^{\mathrm{FA}} = 1/2$ and $\boldsymbol{\mu}_i^V = 0$.

Relation to layer calibration. The visual error in Eq. (7) is the norm of the complete vector in Eq. (41). Thus the visual terms in $\mathcal{E}_l$ and $\mathcal{U}_l$ already account for both omission and compression; no omission penalty is added again. These are single-layer statements evaluated at fixed teacher inputs, with no assumption that all-key attention is preserved inside the visual mixer. They do not bound error propagation after simultaneous replacement of multiple layers.

C.4 HLWA Layer-Error Decomposition

Proof of Proposition 3. For a text query, $\mathcal{P}_i^t = \mathcal{A}_i \cap \mathcal{T}$ contains all legal text keys, and $\mathcal{R}_i = \mathcal{A}_i \cap \mathcal{V}$ contains the omitted visual keys. The retained full-attention mass is $1-\tau_i$, where τ_i is the attention mass on those visual keys. Since the text branch uses the same logits and only renormalizes over $\mathcal{P}_i^t$,

$$\mathbf{o}_i^{\mathrm{text}} = \frac{1}{1-\tau_i}\sum_{j\in\mathcal{P}_i^t} p_{ij}\mathbf{v}_j = \boldsymbol{\mu}_i^P. \tag{49}$$

Partitioning full attention over $\mathcal{P}_i^t$ and $\mathcal{R}_i$ gives

$$\mathbf{o}_i^{\mathrm{FA}} = (1-\tau_i)\boldsymbol{\mu}_i^P + \tau_i\boldsymbol{\mu}_i^R, \tag{50}$$

and therefore

$$\left\|\mathbf{o}_i^{\mathrm{FA}} - \mathbf{o}_i^{\mathrm{text}}\right\|_2^2 = \tau_i^2 \left\|\boldsymbol{\mu}_i^R - \boldsymbol{\mu}_i^P\right\|_2^2 . \tag{51}$$

Both means lie in the convex hull of $\{\mathbf{v}_j : j \in \mathcal{A}_i\}$, so $\|\boldsymbol{\mu}_i^R - \boldsymbol{\mu}_i^P\|_2 \le D_i$. The $\tau_i = 0$ case is defined as zero; for finite logits and a nonempty text-prefix key set, $\tau_i < 1$.

For visual queries, ϵ_i^{LA} is the norm of the full error vector in Eq. (41), including both omission and compression. Reassembly only permutes query rows, and concatenating heads preserves the Frobenius norm. Summing the squared text-row errors and $(\epsilon_i^{\mathrm{LA}})^2$ over disjoint rows and all heads gives the equality in Eq. (12); the text-row diameter bound gives its final inequality.

Finally, submultiplicativity bounds the post-projection Frobenius error by $\|\mathbf{O}_l^{\mathrm{HLWA}} - \mathbf{O}_l^{\mathrm{FA}}\|_F \|\mathbf{W}_o\|_2$; squaring gives the projection statement. □

D Training Details and Evaluation Benchmarks

D.1 Training Hyperparameters

Table 6 summarizes the per-stage hyperparameters. All stages use AdamW (Loshchilov & Hutter, 2019) with $\beta = (0.9, 0.999)$, gradient clipping at 5.0, and a Warmup-Stable-Decay scheduler (10% warmup, 10% decay).

Table 6: **Per-stage training hyperparameters.** N/A denotes no optimizer training in Stage 2.

	Stage 1 MSE distill.	**Stage 2** Layer select.	**Stage 3** KD	**Stage 4** SFT
Batch size	128	N/A	64	64
Steps	20K	N/A	20K	10K
Learning rate	1×10^{-3}	N/A	5×10^{-4}	5×10^{-5}
Weight decay	0.05	N/A	0.05	0.0^*
Max seq. length	4096	4096	8192	32768

Stage 2 (Layer selection). On a held-out calibration split disjoint from all reported test sets, we measure the normalized Stage-1 operator risk and the validation drop caused by each single-layer replacement. Their percentile ranks are combined via Eq. (15) to retain the highest-scoring 25% of layers as full attention.

Attention configuration. The text branch applies softmax attention to all causally available text keys, with $\mathcal{P}_i^t = \mathcal{A}_i \cap \mathcal{T}$ and $n_i^t = |\mathcal{P}_i^t|$. The reported HLWA models use this rule throughout VLM calibration, training, and evaluation, without a fixed text-window cap or window-size search. The text KV cache therefore grows with the text-prefix length; the per-stage maximum sequence lengths in Table 6 are stage-specific input-length limits, not attention-window sizes. Following Stage 2, we replace 75% of the layers with HLWA and retain the remaining 25% as full-attention layers. For the 36-layer 3B model, we set $\lambda = 0.6$ and $k = 9$, retaining full attention in layers $\{0, 2, 3, 7, 14, 15, 17, 20, 29\}$. For the 28-layer 7B model, we likewise set $\lambda = 0.6$ and $k = 7$, retaining full attention in layers $\{0, 2, 5, 11, 13, 15, 22\}$.

D.2 Video-DiT Operator and Evaluation Details

Operator configuration. The Video-DiT extension uses HunyuanVideo-13B and combines bidirectional ReLU linear attention with local exact attention over visual latent tokens. The global branch aggregates all visual tokens using 3D rotary features, while the local branch preserves local interactions between visual latent tokens. This operator is separate from the VLM's modality routing and Gated DeltaNet recurrence; the VLM's 25% retained-FA setting does not define the DiT configuration.

Normalization of the global branch. Let $\mathbf{a}_i = \mathrm{ReLU}(\mathbf{q}_i)$, $\mathbf{b}_j = \mathrm{ReLU}(\mathbf{k}_j)$, and $\mathbf{R}_i = \mathbf{R}_{3\mathrm{D}}(\mathbf{c}_i)$. Expanding Eq. (11) gives

$$\mathbf{o}_i^{\text{bi-LA}} = \sum_j w_{ij}\mathbf{v}_j, \qquad w_{ij} = \frac{\mathbf{a}_i^\top \mathbf{R}_i^\top \mathbf{R}_j \mathbf{b}_j}{\mathbf{a}_i^\top \sum_r \mathbf{b}_r + \varepsilon_{\mathrm{LA}}}. \tag{52}$$

The denominator uses unrotated, nonnegative features, while the numerator preserves relative 3D positional interactions through the rotations. For $\varepsilon_{\mathrm{LA}} > 0$, the denominator is at least $\varepsilon_{\mathrm{LA}}$. The numerator can nevertheless be signed, and the weights need not sum to one. Consequently, this output is not generally a convex combination of the values. A positive stabilizer prevents a zero algebraic denominator but does not by itself control amplification independently of the feature magnitudes. The VLM compression certificates and layer-error decomposition are not approximation guarantees for this operator or its fusion with local attention. The attention stabilizer $\varepsilon_{\mathrm{LA}}$ is distinct from the AdamW epsilon used below.

Training setup. Following the STA-style adaptation protocol (Zhang et al., 2025b), we fine-tune HunyuanVideo-13B on 2,000 synthetic videos generated by its frozen full-attention teacher using Mixkit prompts. Each video contains 117 frames at 1280×768 resolution, with VAE latents and text embeddings precomputed. Training runs for 1,600 steps on eight NVIDIA A800 GPUs with FSDP and context parallelism, using a global batch size of 2, a constant learning rate of 2×10^{-5}, and seed 42. We use AdamW with $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01 excluding biases and normalization parameters, and gradient clipping at 1.0. All student DiT parameters are updated; the VAE, external text encoders, and teacher remain frozen. Computation uses BF16, with linear states, reductions, and losses accumulated in FP32.

Inference and VBench evaluation. The reported sampling setup uses 50 Euler steps, flow shift 7, embedded guidance 6, and external classifier-free guidance (CFG) (Kong et al., 2024). Evaluation generates 129 frames at 1280×720 and 24 fps, corresponding to 118,800 visual tokens, on NVIDIA A800 GPUs. We follow the official 16-dimension VBench protocol (Huang et al., 2024) and `VBench_full_info.json`, generating five samples per prompt and 25 for temporal-flickering evaluation with static-video filtering. The reported comparisons share prompts, seeds, initial noise, and sampling settings, without sample selection or prompt rewriting.

D.3 Evaluation Benchmarks

We evaluate on three groups of benchmarks:

- **Image understanding** (11 benchmarks): MMMU (Yue et al., 2024), MME (Fu et al., 2025a), MMBench (Liu et al., 2024b), MMStar (Chen et al., 2024), AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), OCRBench (Liu et al., 2024c), RealWorldQA.
- **Video understanding** (5 benchmarks): MVBench (Li et al., 2024), LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2025), VideoMME (Fu et al., 2025b), LVBench (Wang et al., 2025a).
- **Video generation**: VBench (Huang et al., 2024) on HunyuanVideo-13B.

E Runtime Evaluation Details and Scope

Profiling configuration. Figure 5 reports measurements at 4K, 8K, 16K, 32K, and 64K tokens. Its end-to-end comparison uses HyLin-7B and Qwen2.5-VL-7B; the default hybrid retains 25% FA layers. At 64K, the measured operator forward/backward ratios are $31.2\times$ /$21.6\times$, prefill/decode ratios are $2.1\times$ /$3.6\times$, and memory reductions are $3.4\times$ and $16.1\times$ for 25% and 0% retained FA, respectively. All operator, prefill, decode, and memory results in the figure were measured using eight NVIDIA A800 GPUs.

Metric definitions. For matched workloads, latency speedup is defined as $t_{\mathrm{FA}}/t_{\mathrm{hybrid}}$ and memory reduction as $M_{\mathrm{FA}}/M_{\mathrm{hybrid}}$. Decode throughput is the number of generated tokens included in the timed interval, summed across the batch, divided by that interval.

Architectural scope of the memory comparison. HLWA maintains a fixed-size visual recurrent state but the complete causal text KV prefix. At fixed dimensions, text-cache storage grows as $\mathcal{O}(N_t)$ and text-attention prefill as $\mathcal{O}(N_t^2)$. Retained FA layers additionally incur $\mathcal{O}(N)$ KV storage and $\mathcal{O}(N^2)$ prefill for $N = N_v + N_t$. Gains therefore depend on modality composition. The 0% FA curve is a resource-profiling ablation, not a quality-validated deployment alternative. Under the described modality-separated attention and token-wise remaining operations, removing every FA layer removes the path from visual-token contents to text outputs, conditional on fixed text inputs and positional/layout information and initially unmixed text embeddings. Shared parameters and residual connections do not restore that missing path.

F Additional Controls and Validation Details

This section complements the training protocol in Appendix D with layer-selection controls, λ sensitivity, and video-frame scaling.

F.1 Layer-Selection Controls

We compare five layer-selection strategies at 25% retained FA. Uniform retains evenly spaced layers, while Bound-only, Realized-only, Benchmark-only, and Joint select the top-k layers using $\mathcal{R}_l^{\text{th}}$, r_l, d_l, and s_l, respectively. Bound-only uses a calibration-dependent upper bound containing measured visual replacement error, and Joint uses $\lambda = 0.6$.

All selectors share the Qwen2.5-VL-3B initialization, calibrated layer candidates, causal attention over the full text prefix without a window cap, and direct-SFT schedule on the same 3M-example subset of LLaVA-OneVision-1.5-Instruct, retaining nine of 36 FA layers. Under this protocol, Joint achieves 77.8 Avg. QA and 54.0 VideoMME (Table 1). Its retained-layer indices are listed in Appendix D.

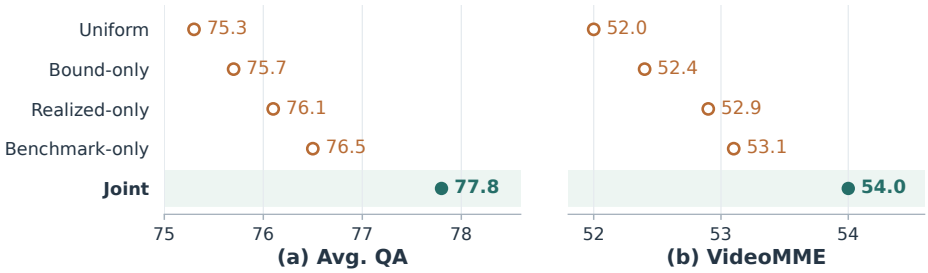


Figure 6: **Layer-selection comparison with 9/36 FA layers.** All selectors use direct SFT on the same 3M examples. (a) Avg. QA across ChartQA, AI2D, and MMBench; (b) VideoMME.

We evaluate $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ under the same protocol (Figure 7). Here, $\lambda = 0$ corresponds to Benchmark-only selection and $\lambda = 1$ to Realized-only selection; intermediate values combine benchmark sensitivity with operator error. The endpoint scores are 76.5/53.1 and 76.1/52.9 in Avg. QA/VideoMME, respectively, consistent with Figure 6. Joint at $\lambda = 0.6$ achieves 77.8/54.0, the highest scores among the six evaluated settings.

F.2 Full-Attention Continued-SFT Control

To distinguish the effects of continued training from those of HLWA, we continue SFT from the released Qwen2.5-VL-3B checkpoint while retaining full attention in all 36 layers. The control uses the same 1.8M-example Stage-4 mixture, next-token prediction objective, data sampling, training budget, optimization settings, and evaluation protocol as HyLin, but excludes attention replacement, layer selection, and the Stage-1 and Stage-3 distillation procedures. As shown in Table 7, continued SFT changes MME from 2157 to 2140 and OCRBench from 797 to 801, while HyLin obtains 2121 and 807, respectively. These results show that continued training accounts for part of the OCRBench improvement but does not improve MME; compared with the continued-SFT control, HyLin further improves OCRBench by 6 points while reducing MME by 19 points. Since HyLin also undergoes

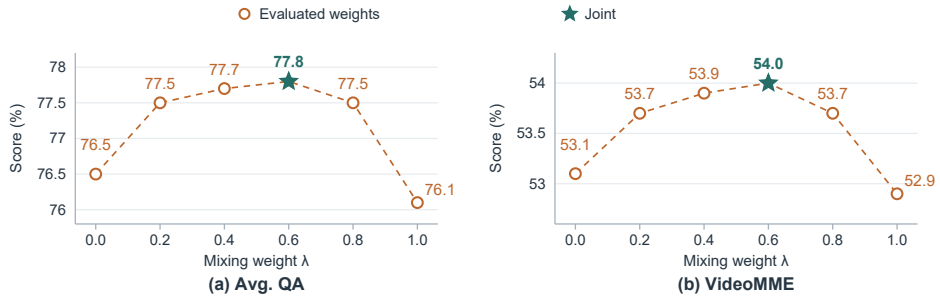


Figure 7: **Sensitivity to λ with 9/36 FA layers.** (a) Avg. QA; (b) VideoMME. Points show measured results at six mixing weights; dashed lines connect the evaluated settings. Stars mark Joint at $\lambda = 0.6$. The endpoints correspond to Benchmark-only ($\lambda = 0$) and Realized-only ($\lambda = 1$). All configurations use direct SFT on the same 3M examples.

Table 7: **Full-attention continued-SFT control at 3B.**

Model	MME ↑	OCRBench ↑
Released FA teacher	2157.0	797
FA teacher + Stage-4 SFT	2140.0	801
HyLin	2121.0	807

the earlier adaptation stages, this comparison provides a matched reference rather than fully isolating the contribution of HLWA.

F.3 Frame Scaling on VideoMME

Figure 8 compares Qwen2.5-VL-3B and HyLin-3B on VideoMME at 8, 64, 128, 256, and 512 input frames. The evaluation scope is VideoMME, not a separately reported long-video subset. The curves show how QA accuracy changes with frame budget; they do not independently establish a single-GPU memory-capacity limit or lossless retrieval of fine-grained visual details.

Execution conditions. The frame-scaling results in Figure 8 were evaluated on eight NVIDIA A800 GPUs. In a separate single-GPU capacity test, Qwen2.5-VL-3B ran out of GPU memory (OOM) at 512 input frames on one NVIDIA A800 GPU. The teacher's 512-frame accuracy point therefore comes from the eight-GPU evaluation; the single-GPU capacity comparison is a separate measurement.

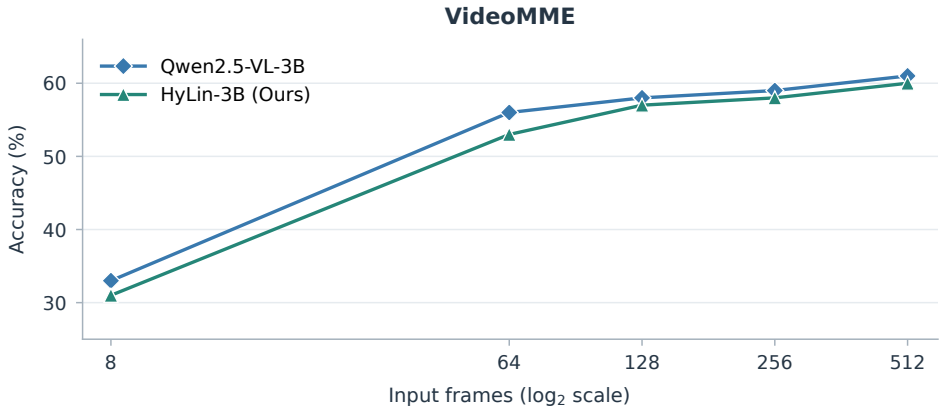


Figure 8: **Frame scaling on VideoMME.** Accuracy of Qwen2.5-VL-3B and HyLin-3B at different input frame budgets, evaluated on eight NVIDIA A800 GPUs. Single-GPU capacity is assessed separately, as described in the text.

G LIMITATIONS AND FUTURE WORK

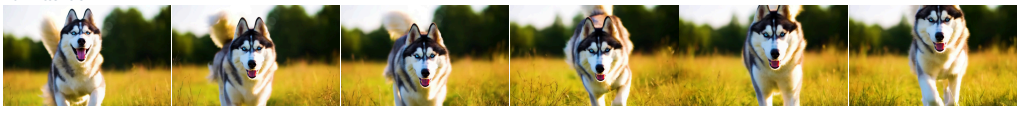
Despite the promising efficiency and task performance of HYLIN, our current evaluation covers a limited range of model scales and training settings. In particular, the VLM experiments focus on adapting pretrained 3B and 7B models, leaving behavior at substantially larger scales and with broader training data to be established. Future work will therefore focus on scaling model capacity and training data, expanding evaluation across more diverse multimodal tasks, and examining how the quality–efficiency trade-off evolves at scale.

H QUALITATIVE VIDEO GENERATION RESULTS

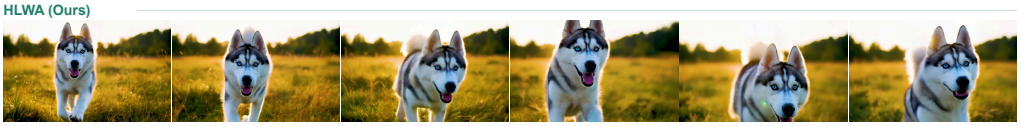
We present qualitative results obtained with HunyuanVideo-13B (Kong et al., 2024) in Figure 9, showcasing the generated video sequences across four distinct prompts describing animal motion.

Prompt: *A fluffy Siberian Husky running joyfully across a sunlit meadow, its pointed ears perked forward and bushy tail raised with excitement. The dog's blue eyes are bright and focused, its mouth slightly open in a playful expression. Its powerful paws bound through the tall grass, creating natural motion blur as its thick gray-and-white fur ripples with each stride. Golden afternoon light filters through the trees in the background, casting soft shadows. The scene radiates warmth and energy. Shot in smooth 4K slow-motion, low-angle close-up tracking shot following the Husky's playful sprint.*

Full Attention

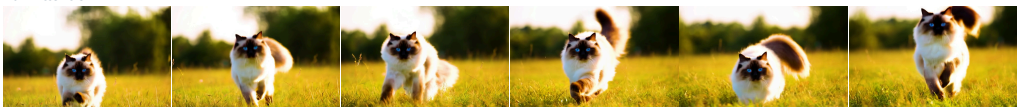


HLWA (Ours)

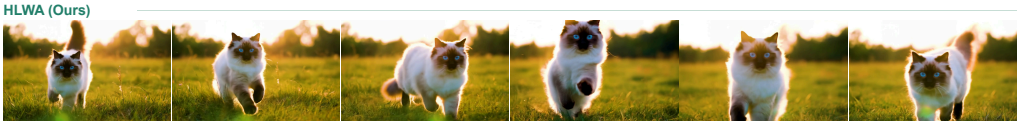


Prompt: *A fluffy Ragdoll cat running joyfully across a sunlit meadow, its ears perked forward and long, feathery tail held high with excitement. The cat's blue eyes are bright and focused, its silky cream-colored fur and darker face, ears, and paws gently illuminated by the sunlight. Its paws swiftly pad through the tall grass, creating natural motion blur as its soft coat flows with each stride. Golden afternoon light filters through the trees in the background, casting soft shadows. The scene radiates warmth and energy. Shot in smooth 4K slow-motion, low-angle close-up tracking shot following the Ragdoll's playful sprint.*

Full Attention

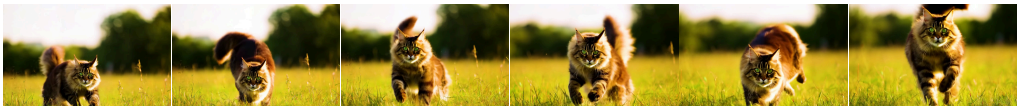


HLWA (Ours)

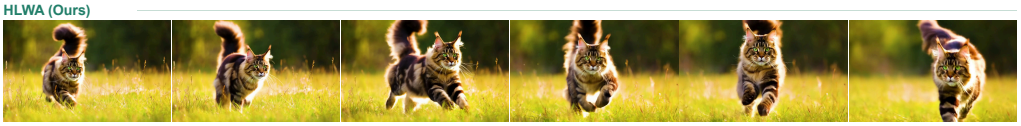


Prompt: *A majestic Maine Coon cat running joyfully across a sunlit meadow, its tufted ears perked forward and long, bushy tail held high with excitement. The cat's green eyes are bright and focused, its shaggy brown tabby coat and flowing chest ruff rippling with every stride. Its large paws bound through the tall grass, creating natural motion blur. Golden afternoon light filters through the trees in the background, casting soft shadows and highlighting individual strands of fur. The scene radiates warmth and energy. Shot in smooth 4K slow-motion, low-angle close-up tracking shot following the Maine Coon's playful sprint.*

Full Attention



HLWA (Ours)



Prompt: *A plush British Shorthair cat running joyfully across a sunlit meadow, its small ears perked forward and thick tail held high with excitement. The cat's round amber eyes are bright and focused, its chubby cheeks and dense blue-gray coat catching the warm sunlight. Its sturdy paws swiftly pad through the tall grass, creating natural motion blur with each playful stride. Golden afternoon light filters through the trees in the background, casting soft shadows. The scene radiates warmth and energy. Shot in smooth 4K slow-motion, low-angle close-up tracking shot following the British Shorthair's playful sprint.*

Full Attention



HLWA (Ours)

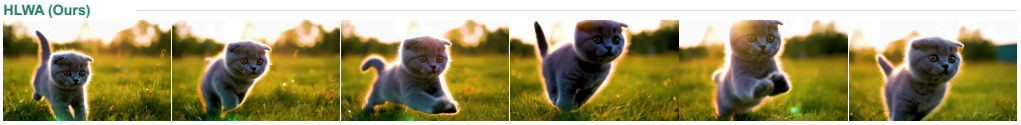


Figure 9: Visual results on HunyuanVideo-13B (Kong et al., 2024) with Full Attention (upper) and HLWA (ours, lower).