Towards Long-window Anchoring in Vision-Language Model Distillation

Haoyi Zhou[1], Shuo Li[2], Tianyu Chen[2], Qi Song[1], Chonghan Gao[2], Jianxin Li[2, 3*]

[1]School of Software, Beihang University, Beijing, China
[2]SKLCCSE, School of Computer Science and Engineering, Beihang University, Beijing, China
[3]Zhongguancun Laboratory, Beijing, China
{haoyi, lishuo2001, tianyuc, songqi23, gaoch, lijx}@buaa.edu.cn

Abstract

While large vision-language models (VLMs) show strong long-context understanding, their prevalent small branches fail on linguistics-photography alignment for a limited window size. We discover that knowledge distillation improves students' capability as a complement to Rotary Position Embeddings (RoPE) on window sizes (anchored from large models). Building on this insight, we propose LAid, which directly aims at the transfer of long-range attention mechanisms through two complementary components: (1) a progressive distance-weighted attention matching that dynamically emphasizes longer position differences during training, and (2) a learnable RoPE response gain modulation that selectively amplifies position sensitivity where needed. Extensive experiments across multiple model families demonstrate that LAid-distilled models achieve up to 3.2× longer effective context windows compared to baseline small models, while maintaining or improving performance on standard Vision-Language benchmarks. Spectral analysis also suggests that LAid successfully preserves crucial low-frequency attention components that conventional methods fail to transfer. Our work not only provides practical techniques for building more efficient long-context VLMs but also offers theoretical insights into how positional understanding emerges and transfers during distillation.

Introduction

The comprehensive understanding and full utilization of long context play a crucial role in building large vision-language models (VLMs). It brings better linguistics-photography alignment in both large scene (Anthropic 2024; Bai et al. 2025; Kamath et al. 2025; Meta 2024; Xue et al. 2024; Zhu et al. 2025) and long storyline (Bai et al. 2025; Meta 2024; Tworkowski et al. 2023; Yu et al. 2024; Zhu et al. 2025), and it helps improving coherence and depth of interaction in multi-rounds dialogue (Anthropic 2024; Bai et al. 2025; Ge et al. 2024; Kamath et al. 2025; Meta 2024; Tworkowski et al. 2023; Xu et al. 2024; Young et al. 2024; Zhu et al. 2025). Currently, large-scale VLMs ($\geq$ 72B parameters) demonstrate the window size scales up to 128k tokens, e.g., Gemma 3 (Kamath et al. 2025), Qwen 2.5-VL (Bai et al. 2025), InternVL 3 (Zhu et al. 2025). To the best of our knowledge, we are the first to point out that the VLMs' prevalent distilled branches ($\leq$ 7B parameters) exhibit markedly constrained window size despite their employment of exact positional embedding, identical architecture, and training methodology. This window shrink is negligible when applying distilled models on short context evaluations, but it becomes a major obstacle during full-length inferencing.

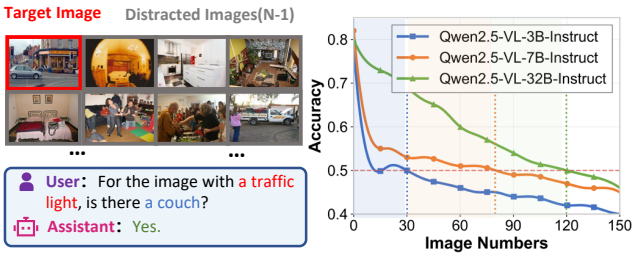

Figure 1: Effective context window comparison. **Left:** The Visual Haystack task requiring retrieval from multi-image inputs. **Right:** Qwen2.5-VL accuracy across scales. Larger models (32B) sustain effective performance (>0.5) significantly longer than smaller counterparts (3B, 7B) despite identical architectures, revealing a scale-dependent RoPE awareness gap that our method targets.

Previous studies have revealed the possibility of extending the pre-trained large language models (LLMs) context window to more than 10 million tokens through various training-stage interventions. Position embedding extrapolation techniques become predominant, e.g., RoPE (Su et al. 2024) enabling models to generalize to sequence lengths beyond their training range, ALiBi (Press, Smith, and Lewis 2022) introducing inductive biases that scale effectively to longer contexts, LongRoPE (Ding et al. 2024) and FoPE (Hua et al. 2024) leveraging the non-uniformity positional interpolation. Besides, fine-tuning approaches on longer texts have shown remarkable effectiveness, as demonstrated in works like LongLLaMA (Tworkowski et al. 2023), which extended context windows through continued pre-training on carefully curated long documents, and Anthropic's Claude model (Anthropic 2024), which achieved 100k token context through specialized training regimes. The above methods mainly focus on the training stage, requiring significant computational resources for model retraining or fine-tuning.

*Corresponding Author.

However, VLMs encounter unique hurdles in handling long windows during training due to their multimodal nature—the visual components introduce substantial complexity in positional understanding across modalities, memory constraints become more severe due to image token density, and the alignment between visual and textual elements at long distances requires fundamentally different mechanisms. More importantly, many foundational assumptions that underpin text-only context extension techniques – such as uniform attention patterns and sequence-independent position encodings – break down when visual tokens with dense, spatially-organized information are introduced into the context window. Few research efforts have thrown light on post-training stage techniques that could extend context windows without expensive retraining.

To better motivate our work, we specialize the concept of ***Long-window Anchoring***. Current state-of-the-art VLMs (Qwen2.5-VL (Bai et al. 2025), InternVL 3 (Zhu et al. 2025), Gemma (Kamath et al. 2025)) often develop models of varying parameter sizes(3B, 7B, 32B) through independent training from scratch, resulting in inconsistent window size capabilities across model sizes. We propose using larger models (e.g., Qwen 2.5-VL 32B) as "anchors" that possess strong long-window capability, then employing post-training methods to align smaller models' long-window capability with these anchors. In this way, smaller models can inherit long-window capability without the prohibitive computational cost of training from scratch, while maintaining their efficiency advantages. Among various potential post-training methods, we begin with knowledge distillation as our fundamental approach – a proven technique for transferring capabilities between models of different sizes while maintaining computational efficiency for the smaller target model.

In this paper, we propose a new perspective to analyze this phenomenon in Figure 1, where we uncover a fundamental distinction: larger VLMs inherently sustain stronger visual haystack performance at extended input image numbers, decaying 5.2× slower compared to 3B models. This positional awareness gap persists even when smaller models demonstrate near-perfect performance on short-context tasks, suggesting the difference stems not from general capacity limitations but specifically from positional representation capabilities. Our analysis reveals that while standard knowledge distillation can unintentionally enhance students' RoPE responsiveness, this effect remains suboptimal without explicit long-context optimization.

To address this challenge, we propose Long-window Anchoring distillation (LAid), a distillation framework that explicitly targets the transfer of long-range attention mechanisms. LAid leverages a Fourier perspective on position distillation through head-level alignment, where each student head learns a weighted combination of multiple teacher heads' query and key representations: $Q^s_{l,i} \approx \sum_{j=1}^{ht} w_{i,j} \cdot Q^t_{L,j}$, $K^s_{l,i} \approx \sum_{j=1}^{ht} w_{i,j} \cdot K^t_{L,j}$. This formulation enables the student to acquire enhanced rotational encoding: $R'\theta(m) = \sum_{j=1}^{ht} w_{i,j} \cdot (W^Q_{t,j} \cdot R_\theta(m) \cdot (W^Q_{t,j})^{-1})$, which expands beyond the frequency limitations of standard RoPE and mitigates frequency leakage in smaller models. Our complete distillation objective combines this position-aware component with traditional knowledge distillation losses, creating a balanced approach that preserves both task performance and positional understanding across contexts of varying lengths.

Our contributions are threefold: (1) We formulate Long-window Anchoring as a distinct problem from traditional context extension—focusing on elevating small models' effective window lengths toward teacher models' upper bounds while preserving computational efficiency; (2) We introduce LAid, a novel distillation framework leveraging head-level alignment with Fourier-enhanced positional knowledge transfer to overcome frequency leakage in compact architectures; (3) We empirically demonstrate the unique challenges vision-language models face with extended contexts—where even 32B-parameter models achieve merely 62.56% accuracy at 100 images, substantially underperforming text-only counterparts. Our approach bridges this capability gap, extending effective context windows by up to 3.2× while preserving overall performance on standard benchmarks.

Related Works

Positional Encoding for Long Sequences. Rotary Position Embeddings (RoPE) and its variants (e.g., M-RoPE (Wang et al. 2024), YaRN (Peng et al. 2024), HiRoPE (Zhang et al. 2024a), 3d-rpe (Ma et al. 2024), Liere (Ostmeier et al. 2024)) have become ubiquitous in modern VLMs due to their simplicity and effectiveness in capturing positional relationships. A common strategy for context length extrapolation involves dynamically adjusting RoPE's rotational frequencies. Position Interpolation (Chen et al. 2023) and YaRN (Peng et al. 2024)apply frequency-based scaling to extend context windows from 4K to over 100K tokens, while LongRoPE (Ding et al. 2024) leverages non-uniform positional interpolation for multi-million token processing. Similarly, ABF (Xiong et al. 2024) achieves 32K context by adjusting base frequency parameters with minimal fine-tuning. Alternative approaches restructure attention mechanisms: SelfExtend (Jin et al. 2024) implements bi-level attention (grouped and neighbor attention) to efficiently handle long-range dependencies without fine-tuning, while DCA (An et al. 2024) decomposes sequences into manageable chunks with specialized attention patterns. MsPoE (Zhang et al. 2024b) and RandomPE (Ruoss et al. 2023)address the "lost-in-the-middle" problem through strategic position encoding modifications. PoSE (Zhu et al. 2024) simulates long inputs by manipulating position indices within short contexts. However, these approaches implicitly assume that positional knowledge transfers uniformly across model scales—overlooking critical variations in RoPE sensitivity between large and small VLMs.

Knowledge Distillation for VLMs. Prior distillation efforts for VLMs have focused on task-specific performance (e.g., LLAVADI (Xu et al. 2024) for VQA, DistillVLM (Fang et al. 2021) for instruction following) or cross-modal alignment (e.g., Align-KD (Feng et al. 2024) for shallow-layer vision-text matching). While recent advances like VLM-KD (Zhang et al. 2024c) address efficiency via text supervision or pruning, and methods such as LongReD (Dong et al. 2025) explore distillation for long-context LLMs with RoPE, none explicitly optimize the long-context ability of VLMs.

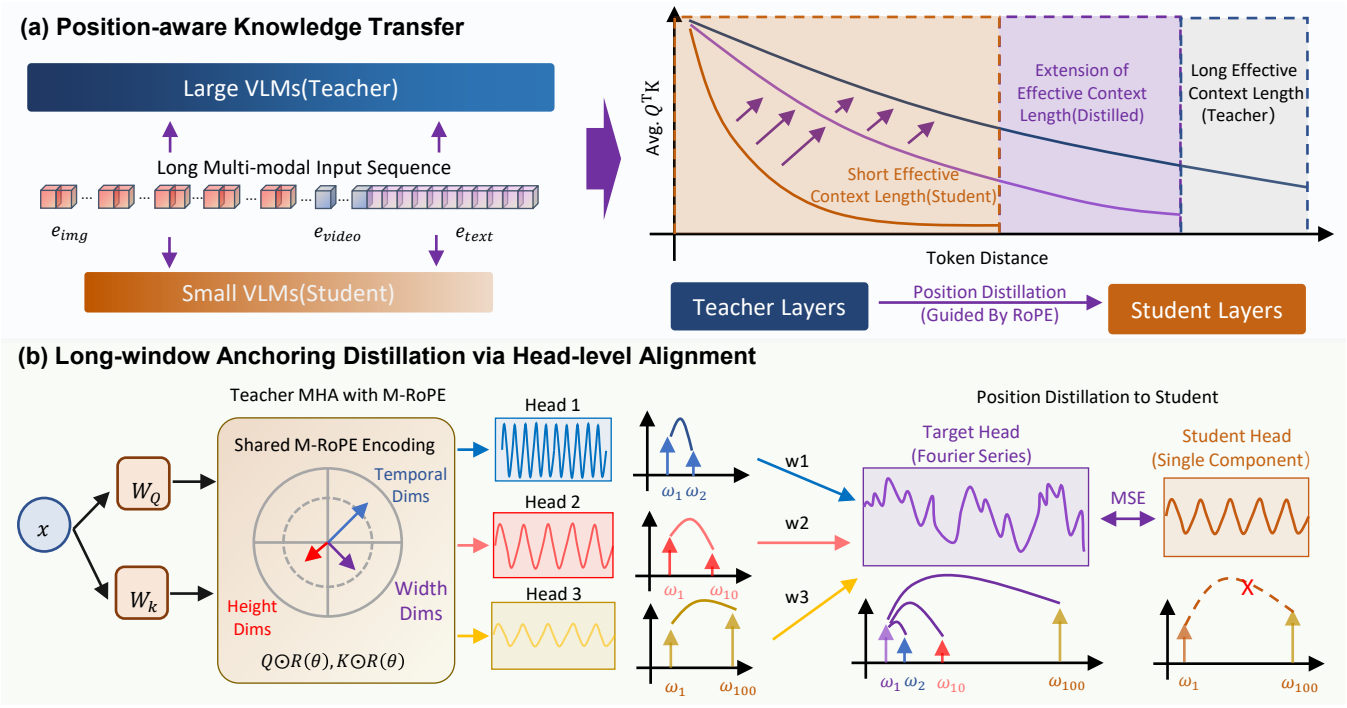

Figure 2: Overview of the LAid framework. **(a) Position-aware Knowledge Transfer:** LAid significantly extends the student's context length (purple) to approach the teacher's capability (gray), far exceeding the baseline (orange). **(b) Fourier-Enhanced Position Distillation:** Teacher attention heads capture positional information across frequency bands via mRoPE. We optimize weights (w) to distill these components into the student head, forming a rich Fourier series representation.

Methods

We formalize our task as follows: given a teacher VLM $\mathcal{T}$ with parameters θ_T (e.g., 32B) demonstrating effective context length L_T, and a student VLM $\mathcal{S}$ with θ_S (e.g., 7B) exhibiting $L_S < L_T$ despite architectural similarity, our goal is to extend $L_S \to L'_S$ where $L'_S \approx L_T$ through post-training alignment. We call it "anchoring" process, which must satisfy two constraints: (1) preserving $\mathcal{S}$'s computational efficiency during inference, and (2) maintaining performance on standard VL benchmarks.

Preliminary

Given a large teacher VLM $\mathcal{T}$ with strong long-context modeling capabilities and a smaller student VLM $\mathcal{S}$, our objective is to enhance $\mathcal{S}$'s long-context performance through position-aware knowledge distillation. We first introduce the key concepts underlying our approach.

Multi-head Attention. The foundation of long-context ability in transformer architectures lies in the attention mechanism, which computes weighted interactions between sequence elements. In self-attention VLMs, these elements can be embedded as sub-word tokens or image patches. The standard attention operation is formulated as: $\text{Attention}(Q, K, V) = \text{softmax}\left(QK^T/\sqrt{d_k}\right)V$, where $Q, K \in \mathbb{R}^{n\times d_k}$ and $V \in \mathbb{R}^{n\times d_v}$ represent the query, key, and value matrices, with n being the sequence length and d_k, d_v the embedding dimensions. Multi-head attention projects these matrices into multiple representation subspaces: $\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$, where $\text{head}_i = \text{Attention}(Q_i, K_i, V_i)$. This structure enables the model to jointly attend to information from different representation subspaces, capturing diverse dependency patterns.

RoPE from a Fourier Perspective. RoPE encodes positional information by applying a rotation matrix to query and key representations. For position m and dimension d, the rotation operation is defined as:

$$\mathbf{q}^{\text{rot}}_{m,d} = R_\theta(m,d)\cdot\mathbf{q}_{m,d},\ \mathbf{k}^{\text{rot}}_{m,d} = R_\theta(m,d)\cdot\mathbf{k}_{m,d}, \quad (1)$$

where:

$$R_\theta(m,d) = \begin{bmatrix}\cos(m\theta_d) & -\sin(m\theta_d)\\ \sin(m\theta_d) & \cos(m\theta_d)\end{bmatrix}, \quad (2)$$

with $\theta_d = 10000^{-2d/D}$ and D being the total embedding dimension. From a signal processing perspective, RoPE encodes positions using a spectrum of frequencies, forming a truncated Fourier series: $f_{\text{RoPE}}(m) = \sum_{d=0}^{D/2-1}[\cos(m\theta_d) + i\sin(m\theta_d)]$. As revealed by recent research (Hua et al. 2024), smaller models suffer more from **frequency leakage and distortion** when handling longer contexts, as their limited capacity constrains their ability to represent the full spectrum of necessary frequencies, leading to rapid attention decay over long distances.

Head-level Position Alignment

As illustrated in Figure 2(b), different attention heads in transformer models capture distinct aspects of contextual

relationships. Research has shown that certain heads (position heads) in large models specialize in modeling long-range dependencies (Hong et al. 2024), while others focus on local interactions. As model size increases, these position heads increasingly dominate the attention distribution, enabling superior long-context modeling.

To transfer this capability to smaller models, we propose a head-level alignment approach. For a student model layer l with head index i and teacher model layer L with heads indexed by $j \in \{1, 2, ..., h_t\}$, we define our distillation objective as learning a set of weights $\{w_{i,j}\}$ such that:

$$Q^s_{l,i} \approx \sum_{j=1}^{h_t} w_{i,j} \cdot Q^t_{L,j}, \quad K^s_{l,i} \approx \sum_{j=1}^{h_t} w_{i,j} \cdot K^t_{L,j}, \tag{3}$$

where $Q^s_{l,i}, K^s_{l,i}$ are the student's query and key matrices, and $Q^t_{L,j}, K^t_{L,j}$ are the teacher's counterparts. This allows each student head to learn from multiple teacher heads, with weights determining the contribution of each teacher head.

Fourier Perspective on Position Distillation

The core insight of our approach comes from viewing the distillation process through Fourier analysis. As shown in Figure 1(b), when RoPE is applied to the teacher model's query and key representations:

$$Q^t_{L,j,\text{rot}} = Q^t_{L,j} \odot R_\theta(m), \quad K^t_{L,j,\text{rot}} = K^t_{L,j} \odot R_\theta(m). \tag{4}$$

Our distillation process learns a linear combination of these frequency-encoded representations:

$$Q^s_{l,i,\text{rot}} \approx \sum_{j=1}^{h_t} w_{i,j} \cdot (Q^t_{L,j} \odot R_\theta(m)). \tag{5}$$

This can be interpreted as learning an enhanced rotational encoding:

$$R'_\theta(m) = \sum_{j=1}^{h_t} w_{i,j} \cdot (W^Q_{t,j} \cdot R_\theta(m) \cdot (W^Q_{t,j})^{-1}), \tag{6}$$

where $W^Q_{t,j}$ represents the teacher's query projection matrix for head j. This formulation enables the student to learn a richer Fourier series representation of positional relationships, expanding beyond the frequency limitations of standard RoPE and mitigating frequency leakage and distortion.

Figure 2(a) illustrates the resultant extension of effective context length achieved through our approach. The student model (purple curve) gains significantly enhanced long-range modeling capabilities compared to its original capacity (orange curve), approaching the performance of the much larger teacher model (gray curve).

Our complete distillation objective is formalized as:

$$\begin{aligned} \mathcal{L}_{\text{LAid}} = \sum_{l,i} \| Q^s_{l,i} - \sum_{j=1}^{h_t} w_{i,j} \cdot Q^t_{L,j} \|^2_F \\ + \| K^s_{l,i} - \sum_{j=1}^{h_t} w_{i,j} \cdot K^t_{L,j} \|^2_F. \end{aligned} \tag{7}$$

This is combined with standard distillation losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{LAid}} \cdot \mathcal{L}_{\text{LAid}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}} + \lambda_{\text{SFT}} \cdot \mathcal{L}_{\text{SFT}}, \tag{8}$$

where $\mathcal{L}_{\text{KL}} = \tau^2 \sum_i p^t_i(\tau) \log \frac{p^t_i(\tau)}{p^s_i(\tau)}$ is the KL-divergence between teacher and student output distributions with temperature τ, where $p^t(\tau)$ and $p^s(\tau)$ are the softened probability distributions of the teacher and student models, respectively, computed as $p^t_i(\tau) = \frac{\exp(z^t_i/\tau)}{\sum_j \exp(z^t_j/\tau)}$ (similarly for $p^s_i(\tau)$), with z^t and z^s being the logits; and $\mathcal{L}_{\text{SFT}}$ is a supervised fine-tuning loss.

Experiments

We address a practical context window extension scenario where only short-context training samples and limited computational resources are available. This constraint prevents straightforward gains from naive data length scaling, necessitating more efficient approaches that maximize the capabilities of student VLMs through architectural innovations or targeted training methods.

Experiments Settings

Baselines. Our comparison includes two categories of approaches: (1) *Length Extrapolation* methods, including YaRN(Yet another RoPE extensioN method) (Peng et al. 2024), which employs frequency-based interpolation strategies for RoPE, and SelfExtend (Jin et al. 2024), which implements bi-level attention (grouped and neighbor attention) to capture both long-range and local dependencies. These methods enable context extension without requiring fine-tuning of pretrained LLMs. In VLMs, we extend them by applying YaRN's RoPE frequency interpolation to textual embeddings for longer text, and potentially to visual positional encodings. SelfExtend's bi-level attention is integrated via an additive mask—based on relative positions within combined visual-textual sequences—into the vision transformer and, critically, cross-modal attention layers, enabling efficient long-context multimodal interaction. (2) *Supervised Fine-Tuning (SFT)*, where we directly fine-tune VLMs via LoRA (Hu et al. 2022) on visual haystack tasks to enhance performance despite limited context.

Models. We evaluated our methodology using the Qwen2.5-VL models (Bai et al. 2025), a prominent family of open-source VLMs. Similar to the previous Qwen2-VL series (Wang et al. 2024), Qwen2.5-VL incorporates Multimodal Rotary Position Embedding (M-RoPE) to effectively fuse positional information across textual, image, and video modalities. During their training phase, these models in 3B, 7B, 32B, and 72B parameter configurations, all support sequence lengths of up to 32,768 Specifically, we conducted a series of experiments utilizing the 7B parameter version as the student model and the 32B parameter version as its teacher model during the distillation progress.

Datasets. We adapt the Visual HayStacks (VHs) benchmark (Wu et al. 2025), which is constructed from the COCO dataset with annotations at the object level (Lin et al. 2014). VHs presents binary (yes/no) questions about anchor and

#Params	Method	Short Window (#img)				Long Window (#img)			Avg. Gain (↑)	
		1	5	10	20	50	100	150	Short	Long
32B	/	83.79	79.11	74.71	73.34	68.17	62.56	60.65	–	–
7B	/	80.22	68.45	62.19	57.21	54.73	51.08	47.43	–	–
	YaRN	80.03	63.78	62.09	55.96	56.26	47.96	42.36	-2.5%	-4.7%
	SelfExtend	78.53	62.58	58.69	53.01	50.35	45.12	40.12	-5.9%	-11.7%
	SFT(LoRA)	**97.78**	**92.92**	**85.80**	**84.73**	63.10	52.28	43.08	**+35.92%**	+3.6%
	LAid (Ours)	92.83	83.26	80.46	74.09	**67.04**	**63.37**	**60.17**	+24.1%	**+24.5%**
3B	/	85.91	65.70	62.09	52.16	50.22	47.80	41.67	–	–
	YaRN	86.27	72.86	57.79	55.74	52.20	45.07	39.97	+2.8%	-1.9%
	SelfExtend	77.89	64.69	54.11	47.95	41.13	35.05	31.25	-7.9%	-23.26%
	SFT(LoRA)	**98.20**	**91.88**	**87.48**	**68.89**	52.34	48.01	42.18	**+31.5%**	+1.96%
	LAid (Ours)	96.83	83.34	74.29	63.27	**58.2**	**53.91**	**50.23**	+20.1%	**+16.4%**

Markers: (1) **Bold** = the best performance per model size; (2) **Underline** = the second performance per model size; (3) LAid = our proposed method; (4) ±% = relative change compared to Base. (5) A vocabulary size mismatch between our 3B parameter model (151936) and the 32B teacher model (152064) renders the KL divergence loss $\mathcal{L}_{\text{KL}}$ incompatible. Consequently, the $\mathcal{L}_{\text{KL}}$ is excluded from the overall training objective $\mathcal{L}_{\text{total}}$. Overall, higher accuracy indicates better performance. (6) The rank of LoRA is 8

Table 1: Performance comparison of different context window extension methods for Vision-Language Models (VLMs) on Visual HayStack benchmark. Evaluated models include 32B/7B/3B parameter versions with five methods: Base (original), YaRN (RoPE extension), SelfExtend, SFT (supervised fine-tuning), and our LAid.

#Images	Avg. Tokens
1	393.46
5	1849.58
10	3630.64
20	7401.08
50	18237.85
100	35418.29
150	53653.22

Table 2: Average input tokens by image count.

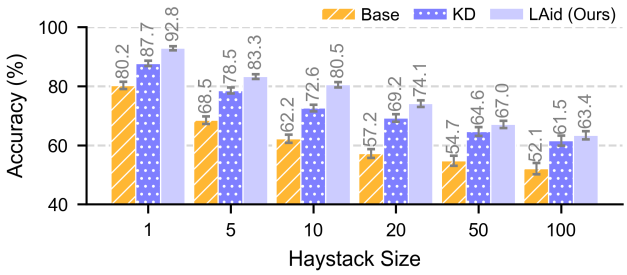

Figure 3: Distillation performance comparison on Visual HayStack dataset. The bar chart illustrates the accuracy of Base, Knowledge Distillation (KD), and our LAid method across increasing haystack sizes (1 to 100 images). Using Qwen2.5-VL-7B as the student model and Qwen2.5-VL-32B as the teacher, LAid consistently outperforms both baseline and standard KD approaches.

target objects in both single-needle (one relevant image) and multi-needle (multiple relevant images) settings. To create a practical evaluation environment, we construct a training set of 5,000 question-answer pairs with haystack sizes ranging from 2 to 20 images. For evaluation, we use 100 samples for each haystack size in the range [1, 2, 5, 10, 20, 50, 100, 150] as shown in Table 2, enabling systematic assessment of model performance across increasing context lengths.

Implementation Details. We trained two student models, Qwen2.5-VL-7B-Instruct and Qwen2.5-VL-3B-Instruct, via knowledge distillation from the Qwen2.5-VL-32B-Instruct teacher model. Both student models were optimized using AdamW, with distinct learning rates set at 1×10^{-5} for the student model parameters and 1×10^{-4} for the weight coefficients. Training was conducted over 10 epochs on 4 NVIDIA A800 GPUs, employing an effective global batch size of 8, realized through a per-device batch size of 1 combined with 8 gradient accumulation steps. The maximum response sequence length was fixed at 512 tokens during training, and a learning rate warmup ratio of 0.05 was applied. We utilized an alpha value of 0.3 for knowledge distillation, specifically targeting the final layer (layer 27 for the 7B model and layer 35 for the 3B model). This training protocol resulted in total durations of approximately 74 hours for the 7B model and 43 hours for the 3B model.

Main Results

Our experimental evaluation reveals significant differences in how various window extension techniques perform on vision-language models, with results shown in Table 1.

Traditional Window Extension Methods Fail to Transfer to VLMs. While large VLMs demonstrate impressive long-context understanding, their smaller counterparts typically suffer from limited effective window lengths, restricting their utility in real-world applications. Interestingly, we ob-

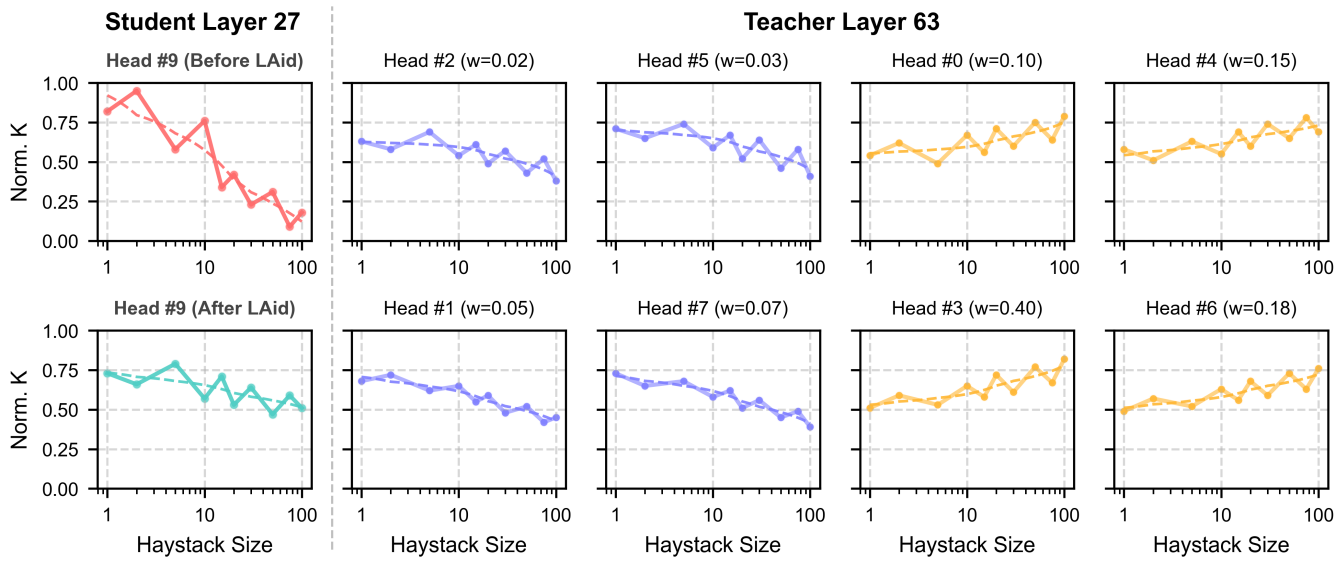

Figure 4: Head-level Knowledge Flow Analysis across different Visual HayStack Size. **The leftmost column** shows the student model's behavior, revealing a dramatic improvement from rapidly decaying activations (before LAid, top) to more stable patterns (after LAid, bottom). **The middle columns** display teachers' Local Position Heads (blue) with a slight downward trend across increasing context lengths, while **the rightmost columns** show Global Position Heads (orange) that maintain or slightly increase activation values at longer distances. LAid transfers this balanced position awareness to the student model, enabling it to maintain activation strength across the full range of context windows rather than focusing primarily on short-range dependencies.

serve that direct application of traditional context extension methods (YaRN, SelfExtend) yields suboptimal results on VLMs. YaRN shows a 4.7% performance degradation on long contexts when applied to Qwen2.5-VL-7B, while SelfExtend performs even worse with an 11.7% decline. This failure to transfer may stem from fundamental differences between unimodal and multimodal models. Unlike pure text LLMs, VLMs must maintain coherent cross-modal alignment across extended contexts, where visual positional embeddings interact with textual ones through complex attention patterns. Traditional methods optimize primarily for token-level dependencies without considering the unique spectral properties of multimodal attention, resulting in frequency distortion that particularly affects long-range visual-textual relationships.

Supervised Fine-Tuning Exhibits Short-Context Bias. Supervised fine-tuning (SFT) demonstrates remarkable performance improvements on short contexts (+35.92% for Qwen2.5-VL-7B), but fails to maintain this advantage as context length increases, showing only modest gains (+3.6%) on long contexts. This pattern reveals an inherent limitation of direct optimization approaches. The short-context bias of SFT can be attributed to its optimization objective, which prioritizes immediate performance gains without explicitly targeting the underlying mechanisms of long-range modeling. SFT effectively enhances content-based attention in familiar context ranges but fails to address the fundamental issue of positional attention decay at extended distances. This leads to overfitting on short-context patterns and poor generalization to longer sequences—precisely the opposite of what window anchoring aims to achieve.

LAid Enables Balanced Long-Short Context Performance. In this paper, we discover a surprising emergent property: knowledge distillation significantly enhances students' responsiveness to Rotary Position Embeddings (RoPE) at greater distances, directly correlating with extended context capabilities. Our proposed LAid method systematically exploits this phenomenon, achieving superior results across the full context spectrum. While LAid's improvements on short contexts (+24.1%) are slightly below SFT's, it substantially outperforms all other methods on long contexts (+24.5%), maintaining consistent performance even at extreme context lengths. This balanced capability stems from LAid's position-aware knowledge transfer approach, which preserves crucial spectral characteristics of the teacher model.

Position-Aware Knowledge Distillation

To dissect how position-aware distillation fundamentally differs from traditional approaches, we conduct a controlled comparison using Qwen2.5-VL-7B as student and Qwen2.5-VL-32B as teacher on the Visual HayStack benchmark. Figure 3 presents performance across increasing context lengths (1 to 100 images), comparing our LAid method against both the baseline model and standard knowledge distillation (KD). This decomposition allows us to isolate and quantify the specific contribution of position-aware distillation beyond conventional semantic knowledge transfer. The results reveal a clear pattern where LAid's advantage becomes increasingly pronounced at longer contexts, precisely where effective positional modeling is most critical.

Performance Pattern. Figure 3 demonstrates LAid's consistent advantage over baseline and KD approaches. The performance gap is most significant at haystack size 1 (5.15%

#Params	Method	Short Window (#img)				Long Window (#img)		Avg. Gain (↑)	
		1	5	10	20	50	100	Short	Long
7B	/	80.22	68.45	62.19	57.21	54.73	47.43	–	–
	w/o $\mathcal{L}_{\text{KL}}$	<u>91.26</u>	<u>84.29</u>	<u>75.09</u>	65.97	<u>66.11</u>	<u>62.29</u>	<u>+18.2%</u>	<u>+20.2%</u>
	w/o $\mathcal{L}_{\text{LAid}}$	87.68	78.50	72.57	<u>69.54</u>	64.61	61.50	+15.5%	+18.1%
	LAid (Ours)	**92.83**	**83.26**	**80.46**	**74.09**	**67.04**	**63.37**	**+24.1%**	**+24.5%**

Markers: (1) **Bold** = the best performance per model size; (2) **Underline** = the second performance per model size; (3) LAid = our proposed method; (4) ±% = relative change compared to Base. Overall, higher accuracy indicates better performance.

Table 3: Performance for loss ablations on VHs datasets, showing accuracy vs. number of images for each configuration.

over KD), and even with longer contexts, LAid still achieves a notable 2.43% improvement at size 50. While all methods show performance decline with increasing context length, LAid maintains superior accuracy throughout the range.

Distillation Limitations. Traditional knowledge distillation transfers task-specific knowledge but fails to capture the position-sensitive representations essential for long-context modeling adequately. At haystack size 100, while KD improves over the baseline by 9.42% (61.50% vs. 52.08%), it still falls short of LAid's 63.37%. This confirms our hypothesis that conventional distillation approaches lack explicit mechanisms for transferring positional understanding.

Spectral Enhancement. LAid's superior performance stems from its Fourier-enhanced approach to position encoding. By transferring a richer Fourier series representation through our enhanced distillation objective (Eq. 7), LAid enables smaller models to overcome the frequency leakage and distortion problems that typically plague them. This spectral enhancement is particularly evident in the consistent performance maintained across extended positional offsets.

Balanced Context Modeling. A distinctive advantage of LAid is its simultaneous excellence in both short-context (1-20) and long-context (50-100) settings. Unlike other context extension methods that often sacrifice near-term accuracy for long-range performance, LAid preserves performance across the entire context spectrum. This balanced capability derives from LAid's preservation of both high and low-frequency components in the positional encoding, enabling comprehensive modeling at all distance ranges.

Head-level Knowledge Flow

To elucidate the mechanisms of LAid, Figure 4 visualizes normalized key activation values across attention heads as context length grows. We analyze heads from Teacher Layer #63 (Qwen2.5-VL-32B) and Student Layer #27 (Qwen2.5-VL-7B), monitoring behavior over haystack sizes from 1 to 100 images. This reveals how LAidtransfers positional awareness at the head level.

Head Specialization. The teacher shows distinct specialization: Local Position Heads (blue, middle) exhibit moderate activations (0.4-0.7) declining gradually with context, focusing on nearby tokens. Global Position Heads (orange, right) maintain stable or rising values (0.5-0.8) for long-range dependencies, evident in high-weight heads (e.g., Head #3, w=0.4).

Student Head Learning Pattern. LAid transforms the student: Pre-distillation (top-left), activations decay rapidly from 0.8 to ¡0.2 due to frequency leakage. Post-LAid(bottom-left), they stabilize at 0.5-0.8, emulating a hybrid of teacher local/global behaviors.

Ablation Study

Our objective combines losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{LAid}} \cdot \mathcal{L}_{\text{LAid}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}} + \lambda_{\text{SFT}} \cdot \mathcal{L}_{\text{SFT}}, \quad (9)$$

where $\mathcal{L}_{\text{LAid}}$ aligns positional heads, $\mathcal{L}_{\text{KL}}$ applies KL-divergence, and $\mathcal{L}_{\text{SFT}}$ handles supervised fine-tuning.

We ablated components to evaluate contributions:

1. Full LAid Loss: Includes all components ($\mathcal{L}_{\text{LAid}}$, $\mathcal{L}_{\text{KL}}$, $\mathcal{L}_{\text{SFT}}$).
2. LAid w/o $\mathcal{L}_{\text{LAid}}$: Trained with $\mathcal{L}_{\text{LAid}}$ and $\mathcal{L}_{\text{SFT}}$ only, excluding general knowledge distillation.
3. LAid w/o $\mathcal{L}_{\text{KL}}$: Trained with $\mathcal{L}_{\text{LAid}}$ and $\mathcal{L}_{\text{SFT}}$ only, excluding general knowledge distillation.

Results in Table 3 span 1-100 image contexts. Removing $\mathcal{L}_{\text{LAid}}$ caused major drops (8.6% short-context, 6.4% long-context), confirming its role in long-range transfer via Fourier alignment. Ablating $\mathcal{L}_{\text{KL}}$ yielded minor declines, emphasizing its support for general knowledge. The complete configuration achieved the best performance, demonstrating complementary effects among the loss functions.

In summary, $\mathcal{L}_{\text{LAid}}$ drives context extension, with $\mathcal{L}_{\text{KL}}$ aiding overall robustness; combined, they anchor smaller models to teacher long-window capabilities.

Conclusions

We present Long-window Anchoring distillation (LAid), tackling window-length gaps in VLMs. By enhancing RoPE responsiveness via position-aware distillation and Fourier-based head alignment, LAid preserves key low-frequency components overlooked by prior methods. Experiments show up to 3.2× context window extension, offering insights into positional transfer across scales.

Limitations include focus on attention (not feed-forward networks) and distillation overhead (unaffecting inference). Future work might integrate efficient tuning or combine with retrieval for ultra-long contexts.

Acknowledgments

This work was supported by the grants from the Natural Science Foundation of China (62225202, 62202029), Young Elite Scientists Sponsorship Program by CAST (No.2023QNRC001) and Beijing Natural Science Foundation (L248032). Thanks for the computing infrastructure provided by Beijing Advanced Innovation Center for Big Data and Brain Computing. This work is also sponsored by CAAI-MindSpore Open Fund, developed on OpenI Community. We owe sincere thanks to all authors for their valuable efforts and contributions. Jianxin Li is the corresponding author.

References

An, C.; Huang, F.; Zhang, J.; Gong, S.; Qiu, X.; Zhou, C.; and Kong, L. 2024. Training-Free Long-Context Scaling of Large Language Models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Anthropic. 2024. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-08-01.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL technical report. http://arxiv.org/abs/2502.13923:2502.13923.

Chen, S.; Wong, S.; Chen, L.; and Tian, Y. 2023. Extending Context Window of Large Language Models via Positional Interpolation. *CoRR*, abs/2306.15595.

Ding, Y.; Zhang, L. L.; Zhang, C.; Xu, Y.; Shang, N.; Xu, J.; Yang, F.; and Yang, M. 2024. LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Dong, Z.; Li, J.; Jiang, J.; Xu, M.; Zhao, W. X.; Wang, B.; and Chen, W. 2025. LongReD: Mitigating Short-Text Degradation of Long-Context Large Language Models via Restoration Distillation. *CoRR*, abs/2502.07365.

Fang, Z.; Wang, J.; Hu, X.; Wang, L.; Yang, Y.; and Liu, Z. 2021. Compressing Visual-linguistic Model via Knowledge Distillation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 1408–1418. IEEE.

Feng, Q.; Li, W.; Lin, T.; and Chen, X. 2024. Align-KD: Distilling Cross-Modal Alignment Knowledge for Mobile Vision-Language Model. *CoRR*, abs/2412.01282.

Ge, S.; Zhang, Y.; Liu, L.; Zhang, M.; Han, J.; and Gao, J. 2024. Model Tells You What to Discard: Adaptive KV Cache Compression for LLMs. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Hong, X.; Jiang, C.; Qi, B.; Meng, F.; Yu, M.; Zhou, B.; and Zhou, J. 2024. On the token distance modeling ability of higher RoPE attention dimension. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, 5877–5888. Association for Computational Linguistics.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hua, E.; Jiang, C.; Lv, X.; Zhang, K.; Ding, N.; Sun, Y.; Qi, B.; Fan, Y.; Zhu, X.; and Zhou, B. 2024. Fourier Position Embedding: Enhancing Attention's Periodic Extension for Length Generalization. *CoRR*, abs/2412.17739.

Jin, H.; Han, X.; Yang, J.; Jiang, Z.; Liu, Z.; Chang, C.; Chen, H.; and Hu, X. 2024. LLM Maybe LongLM: Self-Extend LLM Context Window Without Tuning. *CoRR*, abs/2401.01325.

Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; Rouillard, L.; Mesnard, T.; Cideron, G.; Grill, J.; Ramos, S.; Yvinec, E.; Casbon, M.; Pot, E.; Penchev, I.; Liu, G.; Visin, F.; Kenealy, K.; Beyer, L.; Zhai, X.; Tsitsulin, A.; Busa-Fekete, R.; Feng, A.; Sachdeva, N.; Coleman, B.; Gao, Y.; Mustafa, B.; Barr, I.; Parisotto, E.; Tian, D.; Eyal, M.; Cherry, C.; Peter, J.; Sinopalnikov, D.; Bhupatiraju, S.; Agarwal, R.; Kazemi, M.; Malkin, D.; Kumar, R.; Vilar, D.; Brusilovsky, I.; Luo, J.; Steiner, A.; Friesen, A.; Sharma, A.; Sharma, A.; Gilady, A. M.; Goedeckemeyer, A.; Saade, A.; Kolesnikov, A.; Bendebury, A.; Abdagic, A.; Vadi, A.; György, A.; Pinto, A. S.; Das, A.; Bapna, A.; Miech, A.; Yang, A.; Paterson, A.; Shenoy, A.; Chakrabarti, A.; Piot, B.; Wu, B.; Shahriari, B.; Petrini, B.; Chen, C.; Lan, C. L.; Choquette-Choo, C. A.; Carey, C.; Brick, C.; Deutsch, D.; Eisenbud, D.; Cattle, D.; Cheng, D.; Paparas, D.; Sreepathihalli, D. S.; Reid, D.; Tran, D.; Zelle, D.; Noland, E.; Huizenga, E.; Kharitonov, E.; Liu, F.; Amirkhanyan, G.; Cameron, G.; Hashemi, H.; Klimczak-Plucinska, H.; Singh, H.; Mehta, H.; Lehri, H. T.; Hazimeh, H.; Ballantyne, I.; Szpektor, I.; and Nardini, I. 2025. Gemma 3 Technical Report. *CoRR*, abs/2503.19786.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D. J.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, 740–755. Springer.

Ma, X.; Liu, W.; Zhang, P.; and Xu, N. 2024. 3D-RPE: Enhancing Long-Context Modeling Through 3D Rotary Position Encoding. *CoRR*, abs/2406.09897.

Meta. 2024. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-08-01.

Ostmeier, S.; Axelrod, B.; Moseley, M. E.; Chaudhari, A.; and Langlotz, C. P. 2024. LieRE: Generalizing Rotary Position Encodings. *CoRR*, abs/2406.10322.

Peng, B.; Quesnelle, J.; Fan, H.; and Shippole, E. 2024. YaRN: Efficient Context Window Extension of Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Press, O.; Smith, N. A.; and Lewis, M. 2022. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ruoss, A.; Delétang, G.; Genewein, T.; Grau-Moya, J.; Csordás, R.; Bennani, M.; Legg, S.; and Veness, J. 2023. Randomized Positional Encodings Boost Length Generalization of Transformers. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 1889–1903. Association for Computational Linguistics.

Su, J.; Ahmed, M. H. M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568: 127063.

Tworkowski, S.; Staniszewski, K.; Pacek, M.; Wu, Y.; Michalewski, H.; and Milos, P. 2023. Focused Transformer: Contrastive Training for Context Scaling. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *CoRR*, abs/2409.12191.

Wu, T.-H.; Biamby, G.; Quenum, J.; Gupta, R.; Gonzalez, J. E.; Darrell, T.; and Chan, D. M. 2025. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. arXiv:2407.13766.

Xiong, W.; Liu, J.; Molybog, I.; Zhang, H.; Bhargava, P.; Hou, R.; Martin, L.; Rungta, R.; Sankararaman, K. A.; Oguz, B.; Khabsa, M.; Fang, H.; Mehdad, Y.; Narang, S.; Malik, K.; Fan, A.; Bhosale, S.; Edunov, S.; Lewis, M.; Wang, S.; and Ma, H. 2024. Effective Long-Context Scaling of Foundation Models. In Duh, K.; Gómez-Adorno, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, 4643–4663. Association for Computational Linguistics.

Xu, S.; Li, X.; Yuan, H.; Qi, L.; Tong, Y.; and Yang, M. 2024. LLAVADI: What Matters For Multimodal Large Language Models Distillation. *CoRR*, abs/2407.19409.

Xue, F.; Chen, Y.; Li, D.; Hu, Q.; Zhu, L.; Li, X.; Fang, Y.; Tang, H.; Yang, S.; Liu, Z.; He, E.; Yin, H.; Molchanov, P.; Kautz, J.; Fan, L.; Zhu, Y.; Lu, Y.; and Han, S. 2024. LongVILA: Scaling Long-Context Visual Language Models for Long Videos. *CoRR*, abs/2408.10188.

Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; Yu, K.; Liu, P.; Liu, Q.; Yue, S.; Yang, S.; Yang, S.; Yu, T.; Xie, W.; Huang, W.; Hu, X.; Ren, X.; Niu, X.; Nie, P.; Xu, Y.; Liu, Y.; Wang, Y.; Cai, Y.; Gu, Z.; Liu, Z.; and Dai, Z. 2024. Yi: Open Foundation Models by 01.AI. *CoRR*, abs/2403.04652.

Yu, A.; Nigmetov, A.; Morozov, D.; Mahoney, M. W.; and Erichson, N. B. 2024. Robustifying State-space Models for Long Sequences via Approximate Diagonalization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhang, K.; Li, G.; Zhang, H.; and Jin, Z. 2024a. HiRoPE: Length Extrapolation for Code Models Using Hierarchical Position. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 13615–13627. Association for Computational Linguistics.

Zhang, Z.; Chen, R.; Liu, S.; Yao, Z.; Ruwase, O.; Chen, B.; Wu, X.; and Wang, Z. 2024b. Found in the Middle: How Language Models Use Long Contexts Better via Plug-and-Play Positional Encoding. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zhang, Z.; Meyer, G. P.; Lu, Z.; Shrivastava, A.; Ravichandran, A.; and Wolff, E. M. 2024c. VLM-KD: Knowledge Distillation from VLM for Long-Tail Visual Recognition. *CoRR*, abs/2408.16930.

Zhu, D.; Yang, N.; Wang, L.; Song, Y.; Wu, W.; Wei, F.; and Li, S. 2024. PoSE: Efficient Context Window Extension of LLMs via Positional Skip-wise Training. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; Gao, Z.; Cui, E.; Wang, X.; Cao, Y.; Liu, Y.; Wei, X.; Zhang, H.; Wang, H.; Xu, W.; Li, H.; Wang, J.; Deng, N.; Li, S.; He, Y.; Jiang, T.; Luo, J.; Wang, Y.; He, C.; Shi, B.; Zhang, X.; Shao, W.; He, J.; Xiong, Y.; Qu, W.; Sun, P.; Jiao, P.; Lv, H.; Wu, L.; Zhang, K.; Deng, H.; Ge, J.; Chen, K.; Wang, L.; Dou, M.; Lu, L.; Zhu, X.; Lu, T.; Lin, D.; Qiao, Y.; Dai, J.; and Wang, W. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. http://arxiv.org/abs/2504.10479:2504.10479.