Beyond Sequential Distance: Inter-Modal Distance Invariant Position Encoding

Lin Chen[1,2,3*], Bolin Ni[3†], Qi Yang[1,2,3], Zili Wang[1,2], Kun Ding[1†], Ying Wang[1], Houwen Peng[3], Shiming Xiang[1,2]

[1]MAIS, Institute of Automation, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China
[3]Tencent Hunyuan Team, China

Abstract. Despite the remarkable capabilities of Multimodal Large Language Models (MLLMs), they still suffer from visual fading in long-context scenarios. Specifically, the attention to visual tokens diminishes as the text sequence lengthens, leading to text generation detached from visual constraints. We attribute this degradation to the inherent inductive bias of Multimodal RoPE, which penalizes inter-modal attention as the distance between visual and text tokens increases. To address this, we propose inter-modal **D**istance **I**nvariant **P**osition **E**ncoding **(DIPE)**, a simple but effective mechanism that disentangles position encoding based on modality interactions. DIPE retains the natural relative positioning for intra-modal interactions to preserve local structure, while enforcing an anchored perceptual proximity for inter-modal interactions. This strategy effectively mitigates the inter-modal distance-based penalty, ensuring that visual signals remain perceptually consistent regardless of the context length. Experimental results demonstrate that by integrating DIPE with Multimodal RoPE, the model maintains stable visual grounding in long-context scenarios, significantly alleviating visual fading while preserving performance on standard short-context benchmarks. Code is available at https://github.com/lchen1019/DIPE.

Keywords: Multimodal Large Language Models · Position Encoding · Visual Fading

1 Introduction

Multimodal Large Language Models (MLLMs) [5, 32, 58, 64] have significantly advanced the frontier of machine perception by aligning visual features with the linguistic space of large language models [1, 3, 29, 52]. As model capabilities mature, the research focus has shifted from single-turn visual question answering to complex long-context multimodal understanding [15]. In these scenarios, visual signals are no longer inputs requiring only a glance, but are transformed into persistent contexts that must be continuously indexed throughout the generation of hundreds or thousands of tokens.

* Work done during internship at Tencent Hunyuan Team.
† Corresponding author.

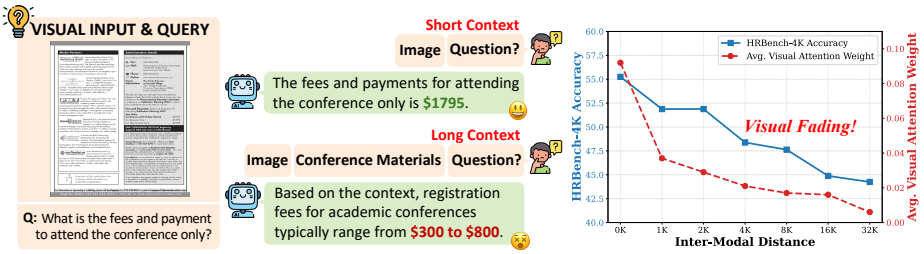

Fig. 1: Illustration of the visual fading phenomenon. **Left**: A qualitative example from DocVQA [40]. While the model accurately grounds visual evidence in a short-context scenario, it fails to preserve correct attention to the image and generates a wrong answer in a long-context one. **Right**: Quantitative analysis of visual fading. As the inter-modal distance increases, the proportion of attention allocated to visual tokens exhibits a sharp decay, indicating that the model gradually looks away from the image.

However, as context length scales, existing MLLMs face a critical limitation, which we term visual fading. As illustrated in Fig. 1 (Left), while MLLMs maintain robust attention to visual inputs and successfully ground visual evidence in a short-context scenario, they fail to attend to correct visual cues and generate erroneous predictions in a long-context one. To systematically quantify this phenomenon, we designed a controlled probing setup that incrementally increases the inter-modal distance between visual inputs and text queries. As shown in Fig. 1 (Right), our empirical results reveal a significant decay in attention allocated to the image as this distance grows. Consequently, the visual constraints progressively erode during text generation, inevitably increasing the likelihood of generating incorrect responses.

While the natural dilution of softmax attention across long contexts partially accounts for this issue, we argue that the inductive bias from position encoding introduces a non-negligible impact. Current MLLMs typically adopt MRoPE [57], which models the spatial and temporal structures of visual and text tokens in a unified sequential framework by decomposing time, height and width components. However, in such a sequential framework, the relative distance between the initial visual tokens and the newly generated text tokens grows monotonically during autoregressive generation. Crucially, MRoPE retains the long-term decay property [47] from the original RoPE, a mechanism designed to impose a distance penalty for capturing linguistic locality. Consequently, this decay imposes a strong inductive bias that visual information becomes increasingly distant as more text is generated, forcing the model to detach from visual constraints. Furthermore, this distance-based penalty stands in contrast to the sustained visual attention of human cognitive mechanisms [19]. Specifically, when a human continuously analyzes an image, the visual reference does not recede into the past like previously spoken words. Instead, it remains right in front of us, regardless of the context length.

To this end, we propose inter-modal **D**istance **I**nvariant **P**osition **E**ncoding **(DIPE)**, a simple but effective method that fundamentally redefines how spa-

tial and temporal geometries are modeled across different modalities. Our core insight is that while intra-modal structural integrity must be preserved, the perceptual distance in inter-modal interactions should remain invariant. Specifically, DIPE orthogonally decomposes the attention mechanism: (1) **Intra-modal Attention**: retains the standard MRoPE to preserve linguistic locality and image's 2D spatial structure. (2) **Inter-modal Attention**: introduces an anchored query that constrains the perceptual distance between text and visual tokens to a constant. By ensuring visual information remains perceptually proximal to the generation process, DIPE effectively mitigates distance-induced visual fading. DIPE can be seamlessly integrated with RoPE variants in MLLMs without introducing additional parameters. Furthermore, it is compatible with modern efficient attention kernels [16] and KV Cache infrastructure.

Experimental results across 19 benchmarks demonstrate that DIPE effectively mitigates the growing visual fading, yielding a 4.10% average accuracy gain over the MRoPE baseline in long-context scenarios. Crucially, these gains are attained without compromising the performance on standard short-context scenarios. Furthermore, in-depth analyses reveal that DIPE restores visual attention in shallow layers, and consistently maintains significantly higher visual attention weights compared to MRoPE as the context length scales. These findings indicate that DIPE provides a principled and effective approach to mitigating visual fading in MLLMs.

2 Related Work

2.1 Multimodal Large Language Models

Advancements in large language models [1, 3, 29, 52] have revolutionized multimodal learning by unifying visual and linguistic representations. To achieve modality alignment, early methods employed cross-attention [2] or Q-Formers [26], while LLaVA [32] later introduced a linear projection approach that has become the mainstream paradigm. Recent works have expanded MLLM capabilities to support dense prediction tasks like segmentation [33, 36] and explore lightweight designs [7, 9]. Numerous state-of-the-art models have been developed, including the GPT series [21] and Gemini [48, 49] series, alongside open-source counterparts such as MiniCPM-V series [64, 65], Qwen-VL series [4–6, 57], InternVL series [10, 14, 58, 66] and LLaVA series [30–32].

2.2 Position Encoding in MLLMs

Since the self-attention mechanism is inherently permutation-invariant, incorporating positional information is indispensable for Transformers to model sequence order [54]. Existing approaches generally fall into two categories: absolute position encoding (APE) [11, 24, 54], which assigns fixed or learnable vectors to specific token indices, and relative position encoding (RPE) [17, 44, 45], which models pairwise distances to capture translation invariance. Rotary Position Embedding (RoPE) [47] has recently emerged as the mainstream foundation for

modern LLMs [3, 52], as it mathematically unifies the benefits of both absolute and relative schemes through rotation matrix while demonstrating superior length extrapolation capabilities [43].

In the context of MLLMs, however, applying standard 1D RoPE to flattened visual tokens disrupts spatial-temporal structures. Consequently, recent studies have extended RoPE to higher dimensions, such as 2D-RoPE [18] for spatial locality in images and 3D MRoPE [57] for MLLMs. Building on these foundations, VideoRoPE [60] and HoPE [25] optimize encoding specifically for video dynamics, V2PE [15] explores variable visual position encoding, Circle-RoPE [55] projects image token indices onto a ring, MRoPE-I [20] revisits frequency allocation strategies and is adopted in Qwen3-VL [5]. However, they still penalize inter-modal interactions over distance and trigger visual fading. In contrast, our DIPE ensures constant inter-modal proximity to reflect sustained visual attention. Importantly, rather than competing with these MRoPE variants, DIPE is fully compatible with them, offering a complementary formulation that resolves visual fading while preserving their sophisticated structures.

3 Preliminaries

3.1 Rotary Position Embedding (RoPE)

RoPE [47] encodes positional information by rotating the query and key vectors within a high-dimensional space. Given a query vector $\boldsymbol{q} \in \mathbb{R}^d$ and a key vector $\boldsymbol{k} \in \mathbb{R}^d$ corresponding to positions m and n, RoPE applies rotation matrices $\boldsymbol{\mathcal{R}}_m \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\mathcal{R}}_n \in \mathbb{R}^{d\times d}$ derived from their absolute position indices. The attention score with RoPE is computed as:

$$\text{Attn}(\boldsymbol{q}, \boldsymbol{k}) = (\boldsymbol{\mathcal{R}}_m \boldsymbol{q})^\top (\boldsymbol{\mathcal{R}}_n \boldsymbol{k}) = \boldsymbol{q}^\top \boldsymbol{\mathcal{R}}_{n-m} \boldsymbol{k}, \tag{1}$$

where $\boldsymbol{\mathcal{R}}_{n-m}$ is a matrix that depends solely on the relative distance $n - m$, thereby naturally incorporating relative positional information into the attention mechanism. The rotation matrix $\boldsymbol{\mathcal{R}}_p$ (where $p \in \{m, n\}$) is constructed as a block-diagonal matrix, where each block rotates a two-dimensional subspace of the embedding. The rotation angle for the j-th subspace is determined by the product of the position index p and a frequency parameter $\theta_j = b^{-2j/d}$, where b is a predefined base constant (e.g., 10000). This enables the long-term decay property as the relative distance increases.

3.2 Long-term Decay

RoPE encodes the relative distance between a query at position m and a key at position n via a rotation matrix $\boldsymbol{\mathcal{R}}_{n-m}$. Su et al. [47] demonstrate that the attention magnitude can be strictly bounded by:

$$|\boldsymbol{q}^\top \boldsymbol{\mathcal{R}}_{n-m} \boldsymbol{k}| \leq \mathcal{C} \sum_{j=1}^{d/2} \left| \sum_{l=1}^{j} e^{i\theta_l (n-m)} \right|, \tag{2}$$

where $\mathcal{C}$ is a bounding constant dependent on the embeddings. As the distance $n - m$ expands, the continuous phase shifts lead to oscillatory cancellation, strictly driving the attention magnitude to decay. This decay injects a strong inductive bias favoring local context, which is essential for linguistic coherence in LLMs. However, it creates a structural conflict in MLLMs. As the response length grows, the relative distance between the newly generated text query and the fixed visual keys increases linearly. Consequently, the attention mechanism mathematically suppresses visual signals as the generation proceeds, causing the model to gradually look away from the image.

3.3 Multimodal RoPE

While Vanilla RoPE effectively models 1D text sequences, it is suboptimal for MLLMs as it neglects the 2D spatial geometry of images. To address this, Multimodal RoPE (MRoPE) [57] decomposes positional information into three components: temporal (t), height (h) and width (w). For the i-th token, MRoPE assigns a position tuple $\boldsymbol{p}_i = (t_i, h_i, w_i)$. Given a query or key vector $\mathbf{x} \in \mathbb{R}^d$, MRoPE splits it into three chunks, $\mathbf{x}^{(t)}, \mathbf{x}^{(h)}, \mathbf{x}^{(w)}$, and applies RoPE independently to each chunk using the corresponding position component:

$$\text{MRoPE}(\mathbf{x}, \boldsymbol{p}_i) = \text{concat}\left(\boldsymbol{\mathcal{R}}_{t_i}\mathbf{x}^{(t)}, \boldsymbol{\mathcal{R}}_{h_i}\mathbf{x}^{(h)}, \boldsymbol{\mathcal{R}}_{w_i}\mathbf{x}^{(w)}\right). \tag{3}$$

The long-term decay property described in Eq. 2 governs each dimension independently. MRoPE employs distinct positional encoding strategies to preserve the structural integrity of different modalities. (1) **For visual tokens:** to maintain spatial structure, the height and width indices (h_i, w_i) correspond to the grid coordinates of the image patch, while the temporal index t_i remains constant for a given frame. (2) **For text tokens:** to model sequential dependency, text tokens utilize a unified index where the spatial and temporal components are identical (i.e., $t_i = h_i = w_i$). Fig. 2 (I) presents an example of MRoPE.

4 Methodology

We introduce inter-modal Distance Invariant Position Encoding (DIPE), a simple but effective mechanism designed to eliminate visual fading by reconfiguring inter-modal geometric relationships. Section 4.1 formulates the implementation of DIPE. Section 4.2 discusses its compatibility with training and inference infrastructures and summarizes the algorithmic implementation.

4.1 Inter-Modal Distance Invariant Position Encoding

Fig. 2 illustrates the implementation of DIPE. The core insight of our approach is to decouple position encoding based on the modality relationship between the query and the key. Rather than applying a sequential position index assignment,

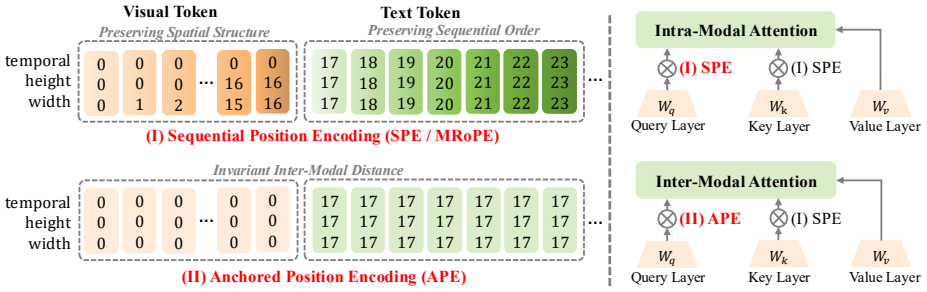

Fig. 2: Overview of inter-modal Distance Invariant Position Encoding (DIPE). DIPE mitigates visual fading by disentangling position encoding based on modality interactions. Specifically, intra-modal attention applies sequential position encoding to both queries and keys to preserve spatial and sequential structures. Conversely, inter-modal attention utilizes anchored position encoding for queries alongside sequential position encoding for keys, effectively anchoring the inter-modal perceptual distance.

we introduce sequential position encoding and anchored position encoding to treat intra-modal and inter-modal interactions orthogonally.

Intra-modal Attention: Preserving Local Structures. Intra-modal attention encompasses interactions where both the query and key originate from the same modality (e.g., text-to-text). In these scenarios, maintaining the fidelity of the relative positions is paramount. Text-to-text attention relies on the distance decay of RoPE to model linguistic locality, whereas visual-to-visual attention demands the preservation of 2D spatial geometry for visual coherence.

To preserve these necessary relative distance properties, we utilize Sequential Position Encoding (SPE), which applies standard MRoPE position tuples, denoted as $\boldsymbol{p}^{\text{spe}}$, to both queries and keys. The intra-modal attention score between a query vector $\boldsymbol{q}$ at position m and a key vector $\boldsymbol{k}$ at position n is computed as:

$$\text{Attn}_{\text{intra}}(\boldsymbol{q}, \boldsymbol{k}) = (\boldsymbol{\mathcal{R}}_{\boldsymbol{p}_m^{\text{spe}}} \boldsymbol{q})^\top (\boldsymbol{\mathcal{R}}_{\boldsymbol{p}_n^{\text{spe}}} \boldsymbol{k}), \tag{4}$$

where $\text{Modality}(\boldsymbol{q}) = \text{Modality}(\boldsymbol{k})$. This formulation strictly retains the structural integrity of both linguistic and visual domains.

Inter-modal Attention: Invariant Inter-Modal Distance. Inter-modal attention specifically refers to queries and keys from different modalities. This is where the distance decay described in Eq. 2 becomes detrimental.

To counteract this, we propose the Anchored Position Encoding (APE). For any inter-modal interaction, we assign anchored position indices to the queries, denoted as $\boldsymbol{p}^{\text{ape}}$. Specifically, for each partitioned modality segment, we extract the position index of its first token. This initial position serves as a unified anchor for all tokens within that entire segment when they act as queries. By applying APE to the query while maintaining SPE for the key, we ensure that the inter-modal relative distance remains invariant with respect to the autoregressive generation steps. The inter-modal attention score between a query vector $\boldsymbol{q}$ at

Algorithm 1 PyTorch-style Pseudocode for Position Design in DIPE.

```
def generate_dipe_indices(
    modality_segments: List[Tensor], # partitioned segments (e.g., [Txt, Img1, Txt, Img2...])
):
    pos_spe, pos_ape = [], []
    seq_offset = 0
    for seg in modality_segments:
        # 1. SPE: adopt indices from MRoPE, and can be replaced with MRoPE variants
        spe_seg = get_mrope_ids(seg, seq_offset)

        # 2. APE: Broadcast the first token's (T,H,W) across the entire segment
        ape_seg = torch.full_like(spe_seg, spe_seg[:, [0]])

        # 3. Update seq_offset and save results for the next segment
        seq_offset = spe_seg.max() + 1 # strategy from MRoPE
        pos_spe.append(spe_seg), pos_ape.append(ape_seg)

    return torch.cat(pos_spe, dim=-1), torch.cat(pos_ape, dim=-1)
```

position m and a key vector $\boldsymbol{k}$ at position n is computed as:

$$\mathrm{Attn}_{\mathrm{inter}}(\boldsymbol{q}, \boldsymbol{k}) = (\boldsymbol{\mathcal{R}}_{\boldsymbol{p}_m^{\mathrm{ape}}} \boldsymbol{q})^\top (\boldsymbol{\mathcal{R}}_{\boldsymbol{p}_n^{\mathrm{spe}}} \boldsymbol{k}), \tag{5}$$

where $\mathrm{Modality}(\boldsymbol{q}) \neq \mathrm{Modality}(\boldsymbol{k})$. This bidirectional anchoring ensures that the perceptual distance between visual and text tokens remains constant.

Implementation of Position Design in DIPE. Algorithm 1 presents the implementation details of the sequential and anchored position index allocation logic in DIPE. For clarity, the input of the function is defined as a list of sequentially partitioned modality segments. SPE naturally adopts the position allocation strategy of MRoPE to maintain local structures. Meanwhile, APE extracts the starting (t, h, w) coordinates of each segment and broadcasts them across all tokens within that segment.

Robustness to Multi-Image and Interleaved Contexts. The formulation of DIPE is inherently designed for interleaved and multi-image contexts. By treating each modality segment as a positional unit, DIPE assigns a unique anchor to each segment based on its relative temporal onset. This unified approach ensures that the macroscopic sequence order is strictly preserved, while the inter-modal perceptual distance for each specific visual entity remains invariant. Experimental results to support this are provided in Sec. 5.3.

4.2 Training and Inference Infrastructure for DIPE

While DIPE theoretically necessitates decoupling the attention computation into intra-modal and inter-modal blocks, we demonstrate that this can be implemented efficiently within modern LLM infrastructures. Our design is compatible with modern efficient attention kernels such as FlexAttention [16] and standard KV cache.

Decoupled Attention with LogSumExp Fusion. Since DIPE applies different rotations to the query depending on the target key, we employ a split-kernel strategy. Specifically, we execute two modality-masked attention kernels.

Algorithm 2 PyTorch-style Pseudocode for Attention With DIPE.

```
def attn_forward_with_dipe(
    hidden_states: Tensor,       # [seq_len, dim]
    pos_spe: Tensor,             # sequential position indices, [3, seq_len]
    pos_ape: Tensor,             # anchored position indices, [3, seq_len]
    intra_attn_mask: Tensor,     # attention mask for intra-modal, [seq_len, seq_len]
    inter_attn_mask: Tensor,     # attention mask for inter-modal, [seq_len, seq_len]
    past_key_values: Optional[Cache] = None,  # KV Cache for inference
):
    # 1. Linear projections of query, key and value
    q, k, v = q_proj(hidden_states), k_proj(hidden_states), v_proj(hidden_states)

    # 2. Apply position encoding
    q_spe = apply_mrope(q, pos_spe)
    q_ape = apply_mrope(q, pos_ape)
    k = apply_mrope(k, pos_spe)

    # 3. Update KV cache if inference
    if past_key_values is not None:
        k, v = past_key_values.update(k, v)

    # 4. Decoupled attention with intra-modal and inter-modal masks
    o_intra, lse_intra = flex_attn_func(q_spe, k, v, intra_attn_mask)
    o_inter, lse_inter = flex_attn_func(q_ape, k, v, inter_attn_mask)

    # 5. Merge via LogSumExp (Eq.6)
    alpha = sigmoid(lse_intra - lse_inter)
    output = alpha * o_intra + (1 - alpha) * o_inter

    return output
```

(1) Inter-modal kernel: computes inter-modal attention using the anchored query and sequential key. (2) Intra-modal kernel: computes intra-modal attention using the sequential query and sequential key.

In our implementation, this split attention is supported by efficient mask-aware kernels such as FlexAttention [16]. To merge the outputs, we utilize the LogSumExp ($\log \sum \exp(\cdot)$) statistic returned by attention kernels. Given outputs $\mathbf{O}_1, \mathbf{O}_2$ and their corresponding LogSumExp normalization factors ℓ_1, ℓ_2, the merged output $\mathbf{O}$ is derived as:

$$\mathbf{O} = \frac{e^{\ell_1}\mathbf{O}_1 + e^{\ell_2}\mathbf{O}_2}{e^{\ell_1} + e^{\ell_2}} = \sigma(\ell_1 - \ell_2)\mathbf{O}_1 + \sigma(\ell_2 - \ell_1)\mathbf{O}_2, \tag{6}$$

where σ is the sigmoid function. The derivation is provided in Appendix 2.

Seamless Integration with KV Cache. DIPE is strictly a query-side intervention. It requires creating two views of the query vector during the forward pass. Critically, it does not alter the content of previous keys and values. This allows DIPE to be deployed on top of pre-computed KV caches without requiring cache re-indexing or memory layout modifications.

Overall Implementation. The execution logic of attention with DIPE is summarized in Algorithm 2. Integrating DIPE into a standard forward pass involves three modifications: First, generating dual query views by applying SPE and APE, respectively. Second, executing attention computations for each query view

using modality-specific masks. Finally, fusing the resulting output vectors via the LogSumExp trick, ensuring mathematically rigorous normalization across the split attention kernels.

5 Experimental Results

5.1 Experimental Settings

Model Architecture. We employ Qwen2.5-3B [50] as our primary LLM backbone, supplemented by Qwen2.5-0.5B and Qwen3-1.7B [63] for scalability analysis. For visual encoding, we primarily utilize SigLIP2-SO400M [53] with the NaFlex variant to support dynamic resolution. The visual features are projected using a simple two-layer MLP with GELU activation.

Training Recipe. Following established paradigms [32], we adopt a two-stage training pipeline. The first stage focuses on vision-language alignment by freezing both the vision encoder and the language model to initialize the visual projector. The second stage performs instruction tuning by unfreezing the LLM. We leverage the LLaVA-Pretrain-558K [32] and LLaVA-NeXT-779K [31] datasets for these two stages, respectively. Detailed configurations are listed in Appendix 1.1.

Evaluation Protocols. To comprehensively assess model performance and its robustness against visual fading, we evaluate our approach on 19 widely adopted benchmarks spanning three core domains: perception, document understanding and general visual question answering (VQA). Specifically, the perception benchmarks include CountBench [41], HRBench [59], V* [61], POPE [28] and BLINK [12]. The document understanding benchmarks consist of ChartQA [38], DocVQA [40], TextVQA [46], AI2D [23], InfoVQA [39] and OCRBench [35]. For general VQA, we evaluate on RealWorldQA [62], MMStar [8], MMBench [34], MMVP [51], MathVision [56] and MathVista [37]. We evaluate these benchmarks under two distinct protocols: Short-Context VQA and Long-Context VQA.

Short-Context VQA protocol strictly follows the standard evaluation settings of the original benchmarks, where the question immediately follows the image. Its primary objective is to verify whether our proposed DIPE preserves the fundamental capabilities of the MLLM without compromising accuracy in standard short-context scenarios.

Long-Context VQA protocol is a controlled diagnostic stress test designed specifically to simulate and quantify the visual fading phenomenon, rather than a conventional long-context benchmark. It systematically expands the relative distance between the visual context and the subsequent task. We source text from the standard distractors corpus [13, 27, 42] and insert them directly between the visual input and the question to simulate long contexts and increase the distance between text and images. We intentionally employ these distractors rather than task-related content to isolate the effect of distance on visual attention, ensuring that the evaluation focuses purely on visual attention degradation without the confounding factors of semantic reasoning over related text. By progressively varying the distractor length from 1K to 32K tokens, this protocol evaluates the model's robustness against visual attention degradation.

Table 1: Performance comparison under the Long-Context VQA protocol. We evaluate Vanilla RoPE [47], MRoPE [57] and MRoPE-I [20] baselines alongside their DIPE-enhanced counterparts.

Capability	Benchmark	Vanilla RoPE		MRoPE		MRoPE-I	
		Base	+DIPE	Base	+DIPE	Base	+DIPE
Perception	HRBench-4K [59]	50.75	51.00(+0.25)	47.63	54.50(+6.87)	51.62	52.12(+0.50)
	HRBench-8K [59]	42.63	44.12(+1.49)	42.25	45.13(+2.88)	43.75	43.88(+0.13)
	V^* [61]	47.12	48.17(+1.05)	49.21	50.26(+1.05)	50.79	47.64(-3.15)
	POPE [28]	78.49	81.37(+2.88)	74.50	85.58(+11.08)	75.13	83.72(+8.59)
	CountBench [41]	70.26	75.15(+4.89)	66.60	75.97(+9.37)	67.62	74.95(+7.33)
	$\text{BLINK}_{\text{val}}$ [12]	40.93	39.19(-1.74)	39.87	41.93(+2.06)	40.29	42.45(+2.16)
Document Understanding	$\text{ChartQA}_{\text{test}}$ [38]	44.20	46.40(+2.20)	40.44	45.16(+4.72)	44.92	47.60(+2.68)
	$\text{DocVQA}_{\text{val}}$ [40]	38.38	40.18(+1.80)	33.25	38.55(+5.30)	36.01	39.90(+3.89)
	$\text{AI2D}_{\text{test}}$ [23]	65.06	66.48(+1.42)	65.32	67.39(+2.07)	66.26	67.39(+1.13)
	$\text{TextVQA}_{\text{val}}$ [46]	48.17	51.60(+3.43)	46.61	50.03(+3.42)	49.75	52.34(+2.59)
	$\text{InfoVQA}_{\text{val}}$ [39]	21.92	21.91(-0.01)	23.60	21.84(-1.76)	22.34	23.26(+0.92)
	OCRBench [35]	23.90	25.20(+1.30)	22.50	23.40(+0.90)	22.50	25.30(+2.80)
General VQA	RealWorldQA [62]	54.51	55.29(+0.78)	52.81	58.82(+6.01)	55.95	56.08(+0.13)
	MMStar [8]	38.53	40.20(+1.67)	38.80	41.67(+2.87)	38.47	40.73(+2.26)
	$\text{MMBenchV1.1-EN}_{\text{dev}}$ [34]	53.64	57.89(+4.25)	50.00	58.75(+8.75)	52.86	56.89(+4.03)
	$\text{MMBenchV1.1-CN}_{\text{dev}}$ [34]	52.01	56.50(+4.49)	49.30	56.66(+7.36)	52.79	54.95(+2.16)
	MMVP [51]	58.67	61.33(+2.66)	56.67	60.00(+3.33)	60.67	61.33(+0.66)
	$\text{MathVision}_{\text{mini}}$ [56]	19.74	22.70(+2.96)	18.09	19.08(+0.99)	19.08	18.75(-0.33)
	$\text{MathVista}_{\text{mini}}$ [37]	30.90	33.20(+2.30)	31.70	32.20(+0.50)	32.70	32.40(-0.30)
Overall	Perception	55.03	56.50(+1.47)	53.34	58.90(+5.56)	54.87	57.46(+2.59)
	Document	40.27	41.96(+1.69)	38.62	41.06(+2.44)	40.30	42.63(+2.33)
	General	44.00	46.73(+2.73)	42.48	46.74(+4.26)	44.65	45.88(+1.23)
	All	46.31	48.31(+2.00)	44.69	48.79(+4.10)	46.50	48.51(+2.01)

Additionally, we employ **MM-NIAH** [22] to evaluate standard long-context capabilities in complex real-world scenarios to complement the Long-Context VQA protocol. This benchmark validates the model's ability to reason over specific needles embedded within extensive interleaved multimodal contexts.

5.2 Main Results

Performance under the Long-Context VQA Protocol. To systematically evaluate robustness against visual fading, we assess the models from two complementary perspectives: a standardized extended-context evaluation (Table 1) and a continuous stress test across scaling context lengths (Fig. 3).

To benchmark the overall capabilities across all 19 datasets, we establish a representative Long-Context VQA setting by introducing 8K tokens of textual distractors. Under this protocol, integrating DIPE consistently yields absolute average improvements of 2.00%, 4.10%, and 2.01% for Vanilla RoPE, MRoPE and MRoPE-I, respectively. While DIPE delivers consistent improvements across the vast majority of tasks, we also observe minor performance regressions on a few specific benchmarks. Nevertheless, considering the consistent overall improvements and the successful prevention of long-term visual fading, these isolated regressions are acceptable.

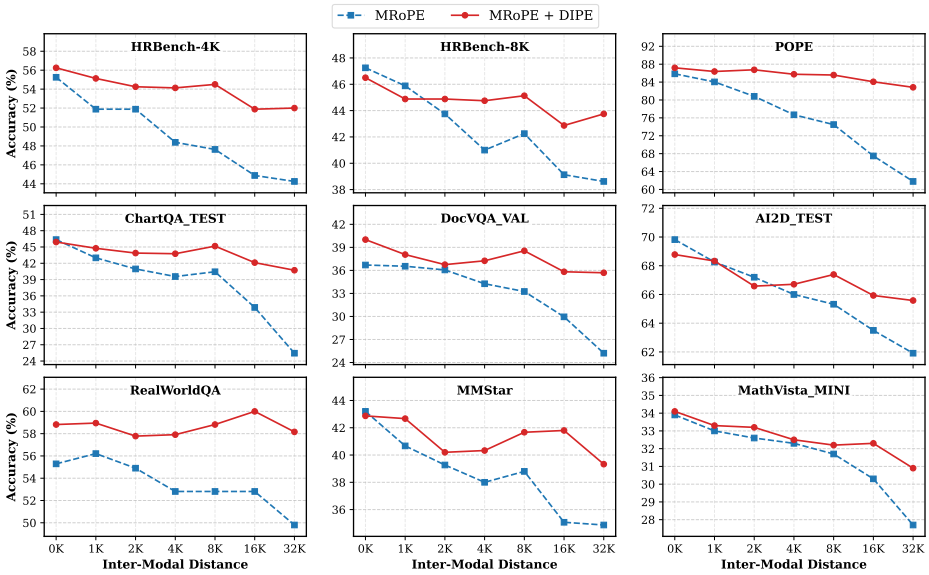

Fig. 3: Accuracy across varying inter-modal distances. We compare the baseline MRoPE against MRoPE + DIPE on 9 benchmarks with distractor lengths ranging from 0K to 32K tokens.

Notably, while the base MRoPE model exhibits lower overall accuracy than both Vanilla RoPE and MRoPE-I, it registers the most substantial gain (+4.10%) when equipped with DIPE. As noted in recent literature [20], this initial baseline degradation stems from MRoPE's suboptimal frequency allocation across spatial dimensions, which inadvertently exacerbates the distance-induced decay in inter-modal attention. By explicitly anchoring the perceptual distance, DIPE effectively neutralizes this structural deficit.

To dynamically assess the progression of visual fading, we evaluate model performance across continuous distractor lengths ranging from 0K to 32K tokens. As illustrated in Fig. 3, the baseline MRoPE suffers a severe and continuous decline in accuracy as the context scales. In contrast, MRoPE + DIPE maintains a remarkably stable performance trajectory. Crucially, the accuracy gap between the two configurations widens progressively as the sequence lengthens. This expanding margin compellingly demonstrates that DIPE effectively neutralizes the escalating inter-modal distance penalty, ensuring persistent visual grounding regardless of the generation length.

Performance under the Short-Context VQA Protocol. While the Long-Context VQA protocol explicitly highlights DIPE's capability to mitigate visual fading, it is equally critical to verify that this architectural modification operates as a non-destructive enhancement in standard short-context multimodal tasks. Fig. 4 presents the results on the Short-Context VQA protocol (the complete results are provided in Appendix 3.1). Quantitative evaluations reveal that DIPE-enhanced models maintain strict performance parity with their baseline counterparts across all three domains. For instance, MRoPE-I equipped with

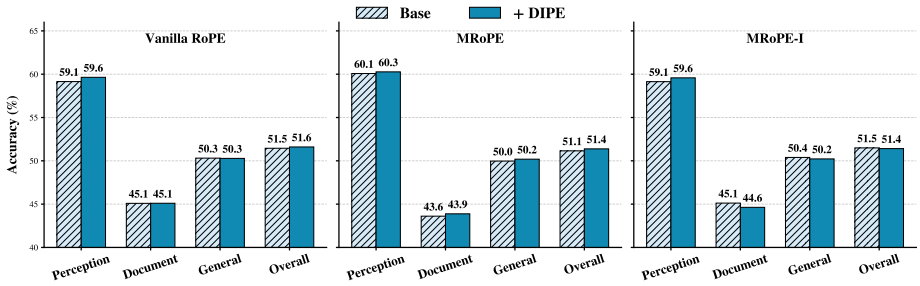

Fig. 4: Performance on the Short-Context VQA protocol. DIPE serves as a non-destructive enhancement, maintaining the performance on standard VQA benchmarks.

DIPE achieves an overall accuracy of 51.4%, effectively matching the 51.5% of the vanilla baseline. This stability remains consistent for both Vanilla RoPE and MRoPE architectures. These results empirically validate our core design principle: by orthogonally decoupling intra-modal and inter-modal geometries, DIPE preserves the essential 2D spatial locality of images and the 1D sequentiality of text in scenarios where visual fading is not yet a factor. This confirms DIPE as a non-destructive enhancement that mitigates visual fading while maintaining strict parity on standard benchmarks.

Performance on the MM-NIAH Benchmark. To further validate the effectiveness of DIPE in standard long-context benchmark, we present the evaluation results on the image reasoning task from the MM-NIAH benchmark in Fig. 5. While both configurations perform comparably at shorter context lengths, the standard MRoPE baseline exhibits a continuous performance degradation as the sequence scales. In contrast, the integration of DIPE effectively mitigates this decay, maintaining a significant performance margin over the baseline. Although both models inevitably experience a sharp accuracy drop at the extreme length of 64K tokens due to the inherent context capacity limits of the underlying language model, DIPE consistently demonstrates superior structural robustness in preserving visual reasoning capabilities.

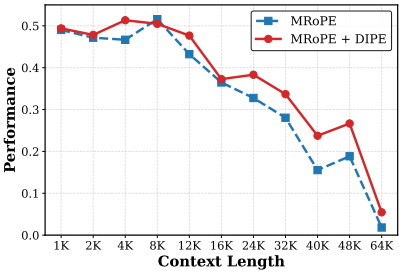

Fig. 5: Performance on the image reasoning task in MM-NIAH [22].

Wall-Clock Efficiency Analysis. We further report wall-clock measurements in Table 2, covering forward, backward and single-token decoding latency. These measurements focus on the attention block, where the additional cost of DIPE is concentrated. DIPE introduces a modest overhead from split attention and LogSumExp

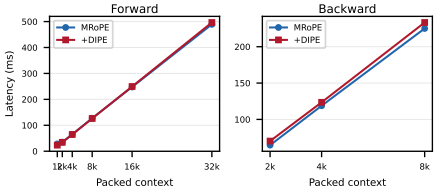

Fig. 6: Training efficiency analysis with sequence packing.

Table 2: Wall-clock latency under unpacked single-sequence evaluation. We report both full-model and attention-block latency.

Full Model (ms)							
Setting	Method	1K	2K	4K	8K	16K	32K
Forward	MRoPE	22.1±0.0	37.2±0.4	68.1±0.4	140.1±0.4	318.5±0.7	818.0±1.6
	+DIPE	28.3±0.5	38.3±0.1	70.6±0.2	144.6±0.5	329.1±0.9	839.8±3.7
Backward	MRoPE	43.0±0.2	69.8±0.4	135.4±0.5	293.0±0.8	OOM	OOM
	+DIPE	47.9±0.5	79.4±0.1	150.6±0.2	322.0±1.0	OOM	OOM
1-token Decoding	MRoPE	18.5±0.1	18.3±0.1	18.1±0.1	18.1±0.3	18.5±0.4	18.3±0.1
	+DIPE	20.1±0.1	20.1±0.1	19.8±0.2	20.1±0.3	20.2±0.1	19.6±0.1
Attention Block (ms)							
Setting	Method	1K	2K	4K	8K	16K	32K
Forward	MRoPE	0.36±0.02	0.41±0.03	0.61±0.01	1.50±0.01	4.23±0.05	13.56±0.01
	+DIPE	0.52±0.02	0.50±0.02	0.70±0.01	1.65±0.00	4.52±0.03	14.31±0.05
Backward	MRoPE	0.83±0.01	1.23±0.02	1.49±0.01	3.62±0.02	10.88±0.03	37.29±0.85
	+DIPE	1.52±0.02	1.54±0.03	2.19±0.01	4.51±0.03	12.46±0.06	40.22±0.34
1-token Decoding	MRoPE	0.24±0.01	0.23±0.01	0.24±0.03	0.23±0.01	0.25±0.01	0.30±0.01
	+DIPE	0.27±0.01	0.28±0.03	0.28±0.01	0.28±0.01	0.31±0.01	0.38±0.01

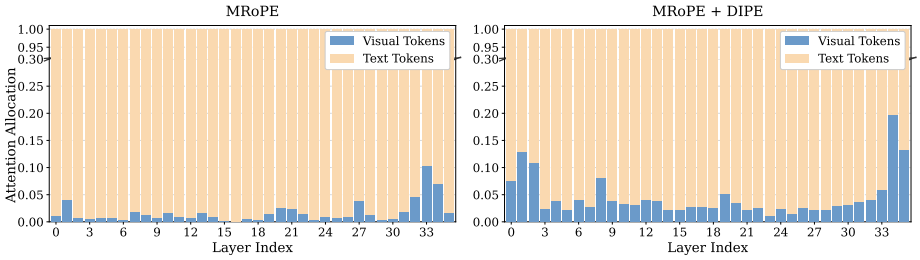

Fig. 7: Layer-wise visual attention allocation analysis. The standard MRoPE baseline exhibits severe attention suppression, particularly in shallow layers. Integrating DIPE effectively restores this distribution across the network.

fusion, while remaining close to the MRoPE baseline across long contexts. Under the common sequence-packing setting, the full-model latency exhibits a similar trend. As shown in Fig. 6, MRoPE and +DIPE nearly overlap for packed training, indicating that the overhead remains small in standard training pipelines.

5.3 In-depth Analysis

Layer-Wise Visual Attention Analysis. To understand the underlying mechanism of how DIPE mitigates visual fading, we visualize the attention weights allocated to visual tokens. Fig. 7 illustrates the layer-wise attention allocation at a fixed distractor length of 8K tokens. Specifically, these weights are computed during the generation phase by averaging the attention scores from query tokens to visual tokens or text tokens across all heads. The standard MRoPE baseline exhibits severe suppression of visual attention, particularly within shallow layers.

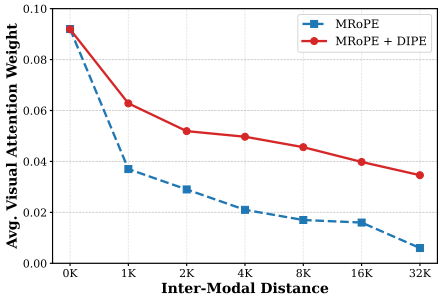

Fig. 8: Average visual attention weight across inter-modal distances.

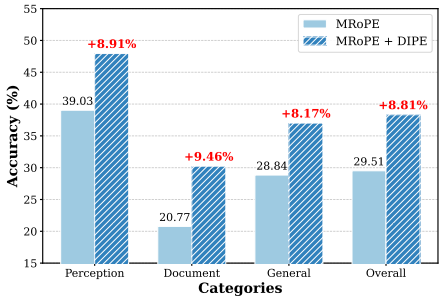

Fig. 9: Performance comparison on Qwen2.5-0.5B as the language model.

It indicates that early layers, heavily governed by spatial encodings, are highly vulnerable to distance-induced penalties. Although deeper layers retain marginal visual attention driven by semantic feature alignment, the visual signal remains significantly attenuated. By anchoring the inter-modal distance, DIPE restores the attention distribution across the network, rescuing positional perception in early layers and reinforcing visual constraints globally.

Mitigating Global Visual Attention Decay. To investigate the global trend of visual fading, we analyze the average visual attention weights across all layers under varying distractor lengths. As depicted in Fig. 8, the baseline MRoPE demonstrates a rapid, monotonic decay in visual attention as the distractor length scales from 0K to 32K tokens, inherently detaching the model from visual constraints. In contrast, the integration of DIPE significantly mitigates this degradation trajectory. It is worth noting that a residual decline in attention persists even with DIPE. This is a natural phenomenon, since attention is inevitably diluted by the additional text tokens as the sequence length increases. However, by enforcing an invariant perceptual proximity, DIPE sustains substantially higher visual attention weights compared to the baseline, even at extreme inter-modal distances. Ultimately, this preserved visual grounding ensures that the model remains constrained by the image during extended generation, directly accounting for the superior accuracy observed in Long-Context VQA.

Generalization Across Model Architectures. To verify the robustness of DIPE across different architectures, we conduct experiments on Qwen3-1.7B. As shown in Table 3, integrating DIPE yields consistent performance improvements across the evaluated domains. Specifically, the model achieves an absolute overall average gain of 4.25%. These enhancements demonstrate that DIPE is a robust and effective solution for mitigating visual fading across different foundation models. The complete results are presented in Appendix 3.2.

Table 3: Performance comparison on Qwen3-1.7B as the language model.

Capability	Base	+DIPE
Perception	50.11	54.92(+4.81)
Document	35.34	36.56(+1.22)
General	35.90	42.25(+6.35)
All	40.21	44.46(+4.25)

Generalization Across Model Scales. To investigate the generalization of DIPE across different scales, we evaluate it on the lightweight Qwen2.5-0.5B model. As shown in Fig. 9, integrating DIPE delivers significant improvements across all domains, achieving an 8.81% increase in overall accuracy. Notably, this gain is even larger than that observed in larger models. We attribute this to the fact that lightweight models inherently possess weaker long-context capabilities, rendering them more susceptible to visual fading. Detailed numerical results are provided in Appendix 3.3.

Robustness of DIPE in Image-Text Interleaved Contexts. DIPE naturally accommodates interleaved image-text scenarios. Since the anchored position index is set to the start of each specific modality segment, distinct visual inputs inherently receive different temporal coordinates. This formulation guarantees that the global macroscopic order of the interleaved sequence is preserved. To verify that our decoupled encoding does not disrupt relative temporal positioning, we evaluated it on a custom 1K-sample interleaved test (Table 4). As expected, MRoPE+DIPE achieves performance strictly comparable to the baseline, confirming that DIPE successfully preserves sequential perception without degradation. Appendix 1.3 provides more details about this interleaved test. Furthermore, this robustness is also empirically validated on the MM-NIAH benchmark (Fig. 5), which specifically targets long-context interleaved reasoning. Additionally, it also support multi-image scenarios and yields improvements on the multi-image BLINK benchmark (Table 1).

Table 4: DIPE maintains accuracy on interleaved tests, demonstrating its applicability to the interleaved data.

Model	**Acc (%)**
Qwen2.5-3B-Instruct	74.2
MRoPE	60.1
MRoPE + DIPE	60.4

6 Conclusion

This paper investigates the underlying mechanism of visual fading in MLLMs, revealing that the distance penalty inherent in MRoPE suppresses long-term inter-modal attention. To address this, we propose inter-modal Distance Invariant Position Encoding (DIPE), a simple but effective mechanism that decouples positional assignments based on modality interactions. It applies the sequential encoding to preserve intra-modal structures and the anchored encoding to anchor inter-modal distance. Extensive experiments demonstrate that DIPE effectively mitigates distance-induced visual fading in long-context scenarios, while preserving short-context accuracy. We hope our insights into inter-modal distance invariance inspire more robust position encoding designs for future MLLMs.

Acknowledgements

This document is the results of the research project funded by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB0500103) and the National Natural Science Foundations of China (Grant No. 62306310). We would like to thank Jiazhen Liu for many fruitful discussions on the early formulation of our method and experimental design.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
6. 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., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
7. Cai, Y., Zhang, J., He, H., He, X., Tong, A., Gan, Z., Wang, C., Xue, Z., Liu, Y., Bai, X.: Llava-kd: A framework of distilling multimodal large language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 239–249 (2025)
8. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
9. Chen, L., Zhao, X., Ding, K., Feng, W., Miao, C., Wang, Z., Guo, W., Wang, Y., Zheng, K., Zhang, B., et al.: Beyond next-token alignment: Distilling multimodal large language models via token interactions. arXiv preprint arXiv:2602.09483 (2026)
10. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)

13. Gao, Y., Xiong, Y., Wu, W., Huang, Z., Li, B., Wang, H.: U-niah: Unified rag and llm evaluation for long context needle-in-a-haystack. arXiv preprint arXiv:2503.00353 (2025)
14. Gao, Z., Chen, Z., Cui, E., Ren, Y., Wang, W., Zhu, J., Tian, H., Ye, S., He, J., Zhu, X., et al.: Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. Visual Intelligence **2**(1), 1–17 (2024)
15. Ge, J., Chen, Z., Lin, J., Zhu, J., Liu, X., Dai, J., Zhu, X.: V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21070–21084 (2025)
16. Guessous, D., Liang, Y., Dong, J., He, H.: Flexattention: The flexibility of pytorch with the performance of flashattention. PyTorch Blog (2024), `https://pytorch.org/blog/flexattention/`, accessed Jun. 24, 2026
17. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654 (2020)
18. Heo, B., Park, S., Han, D., Yun, S.: Rotary position embedding for vision transformer. In: European Conference on Computer Vision. pp. 289–305. Springer (2024)
19. Huang, H., Li, R., Zhang, J.: A review of visual sustained attention: neural mechanisms and computational models. PeerJ **11**, e15351 (2023)
20. Huang, J., Liu, X., Song, S., Hou, R., Chang, H., Lin, J., Bai, S.: Revisiting multimodal positional encoding in vision-language models. arXiv preprint arXiv:2510.23095 (2025)
21. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
22. Kamradt, G.: Needle in a haystack - pressure testing llms. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack` (2023), accessed: 2026-02-10
23. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European Conference on Computer Vision. pp. 235–251. Springer (2016)
24. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
25. Li, H., Qin, Y., Ou, B., Xu, L., Xu, R.: Hope: Hybrid of position embedding for long context vision-language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
26. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning. pp. 19730–19742. PMLR (2023)
27. Li, M., Zhang, S., Zhang, T., Duan, H., Liu, Y., Chen, K.: Needlebench: Evaluating llm retrieval and reasoning across varying information densities. arXiv preprint arXiv:2407.11963 (2024)
28. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
29. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
30. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)

31. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems **36**, 34892–34916 (2023)
33. Liu, J., Feng, M., Chen, L.: Better, stronger, faster: Tackling the trilemma in mllm-based segmentation with simultaneous textual mask prediction. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2026)
34. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European Conference on Computer Vision. pp. 216–233. Springer (2024)
35. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12), 220102 (2024)
36. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)
37. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
38. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: 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 (2022)
39. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)
40. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
41. Paiss, R., Ephrat, A., Tov, O., Zada, S., Mosseri, I., Irani, M., Dekel, T.: Teaching clip to count to ten. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3170–3180 (2023)
42. Park, S., Kim, J., Lee, J., Lee, S.: Dnotitia niah: Enhanced multi-needle retrieval testing framework for evaluating llm long-context understanding capabilities. https://github.com/dnotitia/dnotitia-NIAH (2025)
43. Press, O., Smith, N.A., Lewis, M.: Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409 (2021)
44. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
45. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
46. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019)
47. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
48. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

49. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
50. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/
51. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9568–9578 (2024)
52. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
53. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
55. Wang, C., Guo, J., Li, H., Tian, Y., Nie, Y., Xu, C., Han, K.: Circle-rope: Cone-like decoupled rotary positional embedding for large vision-language models. arXiv preprint arXiv:2505.16416 (2025)
56. Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., Zhan, M., Li, H.: Measuring multimodal mathematical reasoning with math-vision dataset. Advances in Neural Information Processing Systems **37**, 95095–95169 (2024)
57. 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., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
58. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
59. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Yu, W., Tao, D.: Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7907–7915 (2025)
60. Wei, X., Liu, X., Zang, Y., Dong, X., Zhang, P., Cao, Y., Tong, J., Duan, H., Guo, Q., Wang, J., et al.: Videorope: What makes for good video rotary position embedding? arXiv preprint arXiv:2502.05173 (2025)
61. Wu, P., Xie, S.: V?: Guided visual search as a core mechanism in multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13084–13094 (2024)
62. xAI: Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v, accessed: 2026-02-10
63. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
64. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
65. Yu, T., Wang, Z., Wang, C., Huang, F., Ma, W., He, Z., Cai, T., Chen, W., Huang, Y., Zhao, Y., et al.: Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. arXiv preprint arXiv:2509.18154 (2025)

66. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)