Multiple Images Distract Large Multimodal Models via Attention Fragmentation

Tingrui Qiao[1], Di Zhao[⋆1], Yuzhuo Li[1], Bo Pang[1], Caroline Walker[1], Chris Cunningham[2], and Yun Sing Koh[1]

[1] University of Auckland, Auckland 1010, New Zealand
[2] Massey University, Wellington 6021, New Zealand

Abstract. While many tasks require reasoning across multiple images, open-source Large Multimodal Models (LMMs) remain unreliable in these settings. We analyze multi-image LMMs and identify a phenomenon we term attention fragmentation: in each image, tokens at similar background locations act as attention sinks, absorbing disproportionate attention. Causal masking further skews this effect, as earlier images accumulate more sink attention than later images. Using an entropy score over per-image attention, we find that visual focus remains highly dispersed across images rather than isolating key evidence. By applying Pinsker's inequality, we establish a theoretical bound showing that this high-entropy dispersion, combined with stronger early-image sinks, strictly reduces the usable non-sink attention available to earlier images, providing a mechanistic link to image order sensitivity. Motivated by this diagnosis, we propose Attention Remasking (AR), a post-training edit that blocks sink keys and opens a sparse set of cross-image links, routing the recovered attention to task-relevant tokens. AR improves accuracy and reduces order sensitivity across multi-image benchmarks, narrowing the gap between open-source LMMs and leading commercial models.

Keywords: Multi-image Understanding · Large Multimodal Models

1 Introduction

Humans easily draw insight from multiple images, whether comparing photos of similar items [20,22], browsing social media posts [31,32], or following visual instructions [17,23]. Motivated by these natural use cases, recent Large Multimodal Models (LMMs) have begun to accept multiple images via visual tokens, enabling reasoning across images rather than treating each in isolation [4,9,12,33,36]. Researchers have introduced multi-image evaluation benchmarks that test skills such as comparison, retrieval, scene and temporal understanding, and description writing [12,14,30,40]. Despite progress in modeling multiple images, recent benchmark evaluations reveal a substantial performance gap between proprietary and open-source models, with even state-of-the-art systems struggling to achieve robust comprehension. For instance, on MMIU [16], our evaluation of Gemini 3

⋆ Corresponding author. Email: di.zhao@auckland.ac.nz

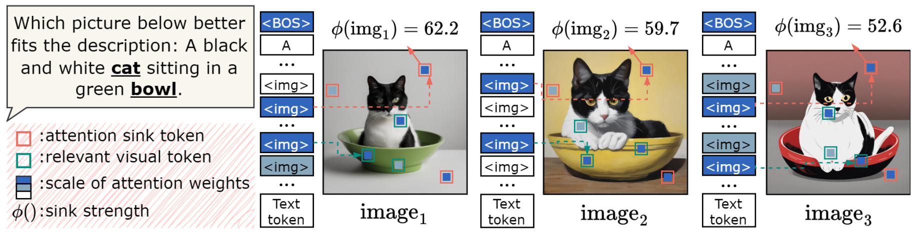


Fig. 1: Attention Fragmentation. In multi-image LMMs, attention concentrates on attention sink tokens recurring in similar background positions in each image, with stronger sinks in earlier images. This diverts attention from task-relevant cues and fragments cross-image focus.

Pro yields only 68% average accuracy, whereas open-source models consistently hover below 56%, indicating that integrating information across images remains a key challenge.

Current multi-image LMMs often interleave visual tokens with text in a single decoder sequence, so each query attends over a mixture of image and text keys under a causal mask [9, 12]. As shown in Fig. 1, we observe a consistent failure pattern in this setting: attention repeatedly concentrates on tokens at similar background locations in each image. These tokens attract a significant amount of attention, distracting the model from task-relevant visual cues. This corresponds to attention sinks, a structural adaptation to softmax normalization that stores excess attention scores without contributing to value computation [7]. Causal masking then makes the effect asymmetric across images: as earlier images are visible to more downstream queries, they accumulate a higher sink attention share than later images. Our entropy-based analysis of per-image attention mass reveals that attention often remains dispersed across images in late layers, rather than shifting toward the image that contains decisive evidence. By applying Pinsker's inequality to bound the deviation from a uniform attention distribution, we mathematically establish that this high-entropy dispersion, combined with stronger early-image sinks, strictly reduces the usable non-sink attention available to earlier images. This provides a mechanistic account of recency bias and image-order sensitivity, in which later images disproportionately influence predictions and answers change with image order [28]. We refer to this compound issue as attention fragmentation, a phenomenon that limits multi-image understanding in current LMMs.

To address this phenomenon, we propose Attention Remasking (AR). AR first masks the identified sink tokens to prevent them from absorbing further attention. To enable cross-image relationships that are otherwise blocked by causal masking, AR then selectively relaxes the mask, permitting forward attention from earlier queries to subsequent visual tokens. To determine exactly which forward links should be opened, we draw on recent mechanistic studies showing that LMMs naturally ground text instructions to visual patches in their earliest layers [37, 38]. Because this shallow-layer attention occurs before deep-layer fragmentation sets in, it provides an uncorrupted map of task-relevant regions. Building on this insight, AR leverages the LMM's own first-layer attention to identify key visual tokens related to the text instruction. We use this native grounding to isolate a sparse set of highly relevant visual tokens across the images. Finally, AR actively routes the attention mass recovered from the blocked sinks directly into these newly unmasked, task-relevant links. By neutralizing sinks and dynamically bridging the causal gap, AR restores cross-image focus and reduces fragmentation.

Our contributions are as follows: (1) We identify attention fragmentation in multi-image LMMs: we demonstrate empirically and theoretically how recurring background attention sinks and causal masking jointly distort attention allocation, providing a mechanistic account of weak cross-image binding and recency bias. (2) We propose Attention Remasking (AR), a lightweight, training-free attention edit. AR blocks sink keys and opens a sparse set of task-relevant cross-image links guided by the LMM's early-layer attention patterns, providing a fully self-contained routing mechanism. (3) Extensive evaluations across multi-image benchmarks show that AR consistently improves accuracy and reduces order sensitivity compared to existing baselines, narrowing the performance gap between open-source LMMs and leading proprietary models.

2 Related Work

Multi-image understanding. Recent large multimodal models process multiple images by interleaving visual and text tokens into a single causal sequence with a shared attention space [6, 9, 12, 15, 27, 33, 41, 42]. To evaluate these capabilities, benchmarks such as MMIU [16], MuirBench [30], and MIRB [40] provide comprehensive coverage of comparison, retrieval, multi-hop reasoning, and spatial-temporal integration tasks. Beyond evaluation, recent training-free methods aim to mitigate multi-image processing failures directly at inference time. For instance, SoFA [28] addresses positional bias by interpolating causal and bidirectional attention via a weighting hyperparameter. Similarly, recent work mitigates cross-image information leakage by scaling the hidden states of delimiter tokens to reinforce intra-image interactions [11].

Attention dynamics in LMMs. Attention sinks are high-activation tokens that absorb surplus attention despite lacking semantic utility [7]. Recent work shows they emerge during LLM pretraining, correlate with massive activations in a few hidden dimensions, and store surplus attention rather than useful signals [7]. In single-image LMMs, sinks often cluster on background patches, prompting post-hoc attention redistribution [10]. Conversely, shallow layers exhibit strong natural grounding, effectively aligning text with relevant visual patches before deep-layer semantic drift occurs [37, 38]. Our work bridges these dynamics in multi-image LMMs: we demonstrate how position-skewed sinks fragment deep-layer attention, and we leverage the uncorrupted early-layer grounding to explicitly route this trapped mass to task-relevant cross-image links.

3 Preliminaries

Attention mechanism. We consider a decoder-based Large Multimodal Model (LMM) with L layers and hidden width D. An input consists of text and M images, tokenized into a single sequence of length N. For each image $m \in \{1,\ldots,M\}$, let $\mathcal{V}_m \subset \{1,\ldots,N\}$ be the indices of its visual tokens, and let $\mathcal{V} = \bigcup_{m=1}^{M} \mathcal{V}_m$ be the set of all visual-token indices. Let $\mathcal{T} \subset \{1,\ldots,N\}$ denote the text-token indices. The j-th input token to the layer ℓ is $x_j^{\ell-1} \in \mathbb{R}^D$, and the stacked states form $X^{\ell-1} \in \mathbb{R}^{N\times D}$. Queries and keys are computed in the standard way, $Q^\ell = X^{\ell-1}W_Q^\ell$ and $K^\ell = X^{\ell-1}W_K^\ell$, with $W_Q^\ell, W_K^\ell \in \mathbb{R}^{D\times d_k}$ and key dimension d_k. The attention score matrix is

$$Z^\ell \;=\; \frac{Q^\ell {K^\ell}^\top}{\sqrt{d_k}} \;+\; \mathcal{M}_{\text{causal}}, \tag{1}$$

where $\mathcal{M}_{\text{causal}} \in \mathbb{R}^{N\times N}$ is the causal mask with $\mathcal{M}_{\text{causal},\,i,j} = -\infty$ for $i < j$ and 0 otherwise. Row-wise softmax yields attention weights

$$\alpha_{i,j}^\ell \;=\; \text{softmax}\big(Z_{i,:}^\ell\big)_j, \qquad \sum_{j\le i} \alpha_{i,j}^\ell = 1. \tag{2}$$

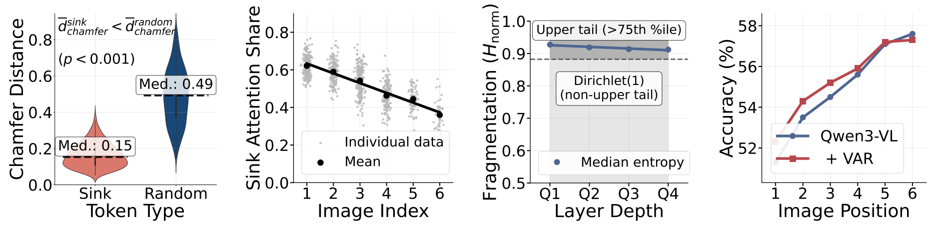
Attention sinks. Prior work reports that Transformer models can allocate large attention to tokens with little semantic value, a behavior termed attention sink, and attributes it to unusually large activations in a small set of hidden dimensions together with the softmax normalization [7]. In LLMs, sinks often occur at fixed positions such as the first token; in LMMs, they appear on visual background tokens [10]. To identify visual sink tokens, we follow the dimension-based criterion used in previous literature [7, 10]. Let $\mathcal{D}_{\text{sink}} \subset \{1,\ldots,D\}$ be the indices of potential sink dimensions. We estimate per-dimension statistics by passing the calibration corpus to the LLM, for each $d \in \mathcal{D}_{\text{sink}}$; let μ_d and σ_d denote the mean and standard deviation of the d-th hidden coordinate measured on the calibration corpus [26]. For any hidden state $x \in \mathbb{R}^D$, define the sink score

$$\phi(x) \;=\; \max_{d\in\mathcal{D}_{\text{sink}}} \frac{x[d]-\mu_d}{\sigma_d}. \tag{3}$$

A token is flagged as a sink at layer ℓ if its previous-layer state satisfies $\phi\big(x_j^{\ell-1}\big) \ge \tau$, where τ is the threshold used in sink-identification work. For image m, the set of visual sink tokens at layer ℓ is

$$\mathcal{S}_m^\ell \;=\; \Big\{\, j \in \mathcal{V}_m \;:\; \phi\big(x_j^{\ell-1}\big) \ge \tau \,\Big\}, \qquad \mathcal{S}^\ell \;=\; \bigcup_{m=1}^{M} \mathcal{S}_m^\ell. \tag{4}$$

This procedure isolates tokens that attract high attention due to massive activation in sink dimensions while contributing little to value computation, as documented for both text and visual sinks.



(a) (b) (c) (d)

Fig. 2: Empirical analysis of attention fragmentation. (a) *Repeated sinks across images:* symmetric Chamfer distance (Eqs. (5) and (6)) between per-image sink sets vs. random baseline. (b) *Positional skew:* sink attention share s_m by image index m (Eq. (7)); earlier images absorb more sink mass. (c) *Level of fragmentation:* normalized entropy H_{norm} across depth (Eq. (8)); values remain high relative to a Dirichlet(1) null, indicating persistent fragmentation. (d) *Positional bias:* position-wise accuracy under shows recency bias and persists after visual attention distribution (VAR).

4 Attention Fragmentation

In Large Multimodal Models (LMMs), multi-image inputs are processed in a single, causally masked sequence, and attention weights are computed via a row-wise softmax over keys [9,12]. Prior work shows that softmax normalization and optimization dynamics can induce attention sinks, tokens that absorb excess attention while contributing little to value computation [7]. Building on this, we empirically examine how attention is allocated across images within the same autoregressive context in LMMs.

4.1 Empirical Observation of Attention Fragmentation

Repeated visual sinks at matched background positions. In multi-image inputs, we observe that per-image sink tokens $\mathcal{S}_m^\ell$ recur in similar spatial regions across images within the same example, drawing substantial attention despite low semantic utility. To quantify the effect, we represent each visual token on the 2-D ViT patch grid of size $R \times C$: token j at row–column (r, c) is mapped to normalized image coordinates $\mathbf{u}_j = (r/R,\ c/C) \in [0,1]^2$; distances are therefore purely spatial in image coordinates. For any unordered image pair (m, m') in the same multi-image example with sink sets $\mathcal{S}_m^\ell = \{\mathbf{u}\}$ and $\mathcal{S}_{m'}^\ell = \{\mathbf{u}'\}$, we measure region-level repetition using the *symmetric Chamfer distance* on the 2-D token locations,

$$d_{\text{Chamfer}}(\mathcal{S}_m^\ell, \mathcal{S}_{m'}^\ell) = \frac{1}{|\mathcal{S}_m^\ell|} \sum_{\mathbf{u} \in \mathcal{S}_m^\ell} \min_{\mathbf{u}' \in \mathcal{S}_{m'}^\ell} \|\mathbf{u} - \mathbf{u}'\|_2 + \frac{1}{|\mathcal{S}_{m'}^\ell|} \sum_{\mathbf{u}' \in \mathcal{S}_{m'}^\ell} \min_{\mathbf{u} \in \mathcal{S}_m^\ell} \|\mathbf{u}' - \mathbf{u}\|_2, \tag{5}$$

where lower values indicate stronger spatial recurrence. For an example with M images, we summarize recurrence by the *per-example average* over all unordered pairs

$$\overline{d}^{\ell}_{\text{Chamfer}} = \frac{2}{M(M-1)} \sum_{1 \leq m < m' \leq M} d_{\text{Chamfer}}(\mathcal{S}^{\ell}_{m}, \mathcal{S}^{\ell}_{m'}). \tag{6}$$

As shown in Fig. 2a, the per-example average $\overline{d}^{\ell}_{\text{Chamfer}}$ across images in the same example is significantly smaller than a random sampled baseline, as confirmed by a *paired Wilcoxon signed-rank test*. For the baseline, for each image, we uniformly sample $|\mathcal{S}^{\ell}_{m}|$ patch tokens at random locations, recompute the distances, and average over repeated draws. This confirms that sink tokens recur in background regions rather than appearing at arbitrary locations.

Masking sinks leaves predictions unchanged. We remove all incoming attention to visual sinks by editing the attention mask at inference: for each layer ℓ and for every query token i, we set $Z^{\ell}_{i,j} = -\infty$ for each key $j \in \mathcal{S}^{\ell}$, which forces $\alpha^{\ell}_{i,j} = 0$ for all queries and all layers. We then compare the model's original outputs with the masked run on the same inputs using the *answer-flip rate*, defined as the fraction of examples whose predicted answer string changes after masking. We report this rate with a *Wilson 95% confidence interval* for binomial proportions; when no flips are observed, we state a conservative 95% upper bound on the true flip probability using the *rule of three*, *i.e.*, $3/n$ for n evaluated items, which closely matches the one-sided Clopper–Pearson bound in this case. Across models, flip rates remain within tight Wilson intervals near zero, indicating that eliminating attention to sink keys does not materially alter predictions. This outcome is consistent with prior reports that attention sinks behave as surplus-attention anchors whose removal has minimal effect on outputs [7, 10].

Positional skew toward earlier images. Within the same multi-image example, sink strength is not uniform across images. We quantify per-image sink strength at layer ℓ by the *Sink Attention Share*

$$\zeta^{\ell}_{m} = \frac{\sum_{i \in \mathcal{V}_m} \sum_{j \in \mathcal{S}^{\ell}_{m}} \alpha^{\ell}_{i,j}}{\sum_{i \in \mathcal{V}_m} \sum_{j \in \mathcal{V}_m} \alpha^{\ell}_{i,j}}, \tag{7}$$

where $\alpha^{\ell}_{i,j}$ are attention weights averaged over heads. Intuitively, ζ^{ℓ}_{m} measures what fraction of the attention budget directed to image m is absorbed by its sink tokens rather than informative visual tokens, so larger values indicate more sink attention. As shown in Fig. 2b, earlier images exhibit consistently larger ζ_m than later ones; a paired Wilcoxon signed-rank test comparing the first and last image within each example rejects the null of equal medians, and a within-example regression of ζ_m on the image index m yields a negative, statistically significant slope. This pattern aligns with the one-way access imposed by causal masking [34]: an image that appears earlier is exposed to more downstream queries, and we observe that ζ_m increases with this exposure; moreover, permuting image order within the same example shifts ζ_m toward the images moved earlier, reinforcing the link between positional access and sink accumulation.

4.2 Measuring Attention Fragmentation

Entropy-based measurement of image-level fragmentation. Prior analyses of Transformer attention report layer-dependent patterns, suggesting that attention can become more task-specific deeper in the stack [1, 29]. In multi-image examples, attention allocation may evolve with depth: early layers can distribute mass more broadly, whereas later layers are expected to place relatively more weight on the image that carries decisive evidence rather than maintaining an even spread. For example, a matching-style query such as "Which image shows the red umbrella?", would expect more attention to be allocated to the image with a red umbrella in the late layers rather than spreading evenly across the images. Following previous work that uses entropy to characterize attention dispersion [3, 8], we measure fragmentation at the image level using entropy to test whether attention becomes concentrated or remains dispersed in late layers. For decoder layer ℓ and query token i, we aggregate attention to image m by defining $p_m(i,\ell) = \sum_{j\in\mathcal{V}_m} \alpha^{\ell}_{ij}$, and define the total visual attention mass as $w(i,\ell) = \sum_{j\in\mathcal{V}} \alpha^{\ell}_{ij}$. We then normalize over the visual mass to obtain an image-level distribution $\tilde{p}_m(i,\ell) = \frac{p_m(i,\ell)}{w(i,\ell)}$, so that $\sum_{m=1}^{M} \tilde{p}_m(i,\ell) = 1$ when $w(i,\ell) > 0$. For numerical stability, we use a smoothed image-level distribution with $\varepsilon = 10^{-8}$, $\bar{p}_m(i,\ell) = \frac{\tilde{p}_m(i,\ell)+\varepsilon}{1+M\varepsilon}$. We then compute the normalized Shannon entropy as

$$H_{\text{norm}}(i,\ell) \;=\; \frac{-\sum_{m=1}^{M} \bar{p}_m(i,\ell)\,\log \bar{p}_m(i,\ell)}{\log M}, \tag{8}$$

which lies in $[0, 1]$: values near 0 indicate concentrated focus on one image, while values near 1 indicate a uniform spread across images. In our analysis, higher normalized entropy indicates stronger attention fragmentation at the image level, whereas lower values indicate more concentrated attention. Mechanistically, if each image contains background sink tokens that attract a comparable share of attention, the per-image attention masses $p_m(i,\ell)$ tend toward equality (i.e., $\approx 1/M$), which increases H_{norm}.

Fragmentation persists throughout layers. Following prior work that examines attention patterns across all layers [1, 39], we summarize normalized entropy by depth using quartile bins. We use a Dirichlet(1) compositional reference because it is uniform over the M-simplex and encodes no preference among images. As shown in Fig. 2c, relative to this reference, early-layer entropy already lies in the upper tail of its Monte Carlo distribution, indicating high dispersion compared to an uninformed spread. Entropy shows no meaningful reduction with depth: paired Wilcoxon signed-rank tests between adjacent quartiles are not significant after Holm correction, and per-example Spearman correlations between the layer index and entropy are near zero. By the final quartile, the median entropy remains in the upper tail of the Dirichlet(1) reference, underscoring persistent fragmentation rather than concentrating on relevant images.

Link between high entropy, skewed sinks, and recency bias. We found that fragmentation persists under permutation of image order, and answers flip with a recency bias; images placed later in the sequence disproportionately influence predictions, consistent with one-way access under causal masking and mirroring behavior reported in recent position-bias studies [28, 34]. Let $p_m(i,\ell) = \sum_{j\in\mathcal{V}_m} \alpha^{\ell}_{i,j}$ denote the per-image attention mass and let ζ^{ℓ}_m be the sink share on image m (Eq. (7)), where $\alpha^{\ell}_{i,j}$ are attention weights averaged over heads. Let $w(i,\ell) = \sum_{m=1}^{M} p_m(i,\ell) = \sum_{j\in\mathcal{V}} \alpha^{\ell}_{i,j}$ be the total visual attention mass. Define the non-sink attention mass $r_m(i,\ell) = p_m(i,\ell)\,(1-\zeta^{\ell}_m)$. If visual attention is evenly distributed across images (high entropy conditional on visual mass), so that $p_m(i,\ell) = w(i,\ell)/M$ for all m, then for any images a, b, $r_a(i,\ell) - r_b(i,\ell) = \frac{w(i,\ell)}{M}\,(\zeta^{\ell}_b - \zeta^{\ell}_a)$, hence $\zeta^{\ell}_a > \zeta^{\ell}_b$ implies $r_a(i,\ell) < r_b(i,\ell)$. More generally, if the masses are near-uniform with $\max_m \big|p_m(i,\ell) - w(i,\ell)/M\big| \le \varepsilon$ for all m, then for any a, b, $r_a(i,\ell) - r_b(i,\ell) \ge \frac{w(i,\ell)}{M}\,(\zeta^{\ell}_b - \zeta^{\ell}_a) - 2\varepsilon$. Here ε quantifies deviation from uniformity across images; when the normalized entropy $H_{\text{norm}}(i,\ell)$ of the conditional distribution $\tilde{p}_m(i,\ell) = p_m(i,\ell)/w(i,\ell)$ is high, ε can be bounded via Pinsker's inequality as $\varepsilon \le w(i,\ell)\sqrt{2\log M\,\big(1 - H_{\text{norm}}(i,\ell)\big)}$ (natural logs). Because causal masking increases sink share for earlier images, these relations reduce their non-sink mass and bias decisions toward later images, manifesting as recency bias.

4.3 Limitations of Post-Softmax Redistribution

Post-softmax redistribution from prior work. Prior research on attention sinks in single-image LMMs proposes reallocating the attention mass removed from identified sink tokens to visual non-sink tokens by directly editing the weights $\alpha^{\ell}_{i,j}$ in (2) computed after the causal mask $\mathcal{M}_{\text{causal}}$ in (1) [10]. The reallocation is proportional to existing non-sink token attention weights, so the image-level distribution $p_m(i,\ell) = \sum_{j\in\mathcal{V}_m} \alpha^{\ell}_{i,j}$ inherits the baseline pattern; when non-sink token attention is already fragmented across images and skewed toward earlier ones, the same dispersion and skew are preserved, sometimes reinforced by larger sink budgets on earlier images. Because the mask is not altered, forward links from earlier queries to later images cannot be created, so order sensitivity induced by one-way access remains.

Fragmentation is unchanged after proportional redistribution. We measure image-level dispersion with the normalized entropy $H_{\text{norm}}(i,\ell)$ in (8). Let $H^{\text{base}}_{\text{norm}}$ and $H^{\text{post}}_{\text{norm}}$ denote the scores before and after redistribution. Paired Wilcoxon signed-rank tests find no meaningful difference between $H^{\text{post}}_{\text{norm}}$ and $H^{\text{base}}_{\text{norm}}$, and Hodges–Lehmann estimates of the median change are near zero. Quartile summaries by depth show that the final-quartile median entropy remains in the upper tail of the Dirichlet(1) compositional null both before and after redistribution, indicating that high dispersion persists throughout the layers.

Order sensitivity and recency bias persist. Using the same answer-flip rate and permutation protocol introduced earlier, we observe similar flip frequencies before and after redistribution, indicating that post-softmax reweighting does not stabilize predictions under image reordering. As shown in Fig. 2d, position-

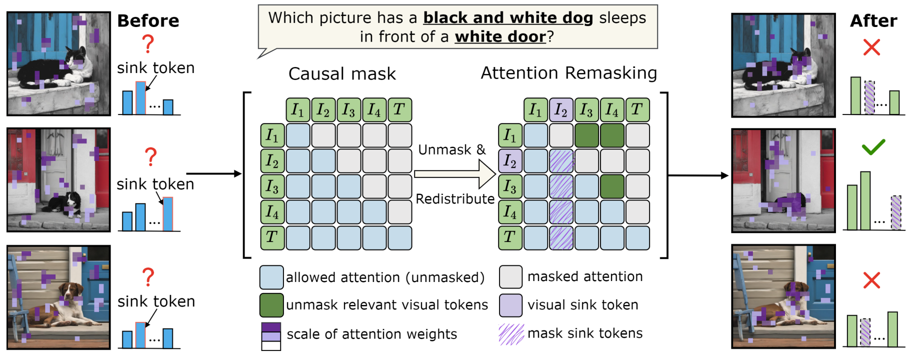


Fig. 3: Attention Remasking. A post-training edit to pre-softmax attention that masks visual sink tokens and selectively unmasks links between visual tokens guided by task relevance. AR reduces attention fragmentation and mitigates order sensitivity in multi-image LMMs.

wise accuracy continues to increase with later positions, consistent with a recency bias arising from one-way access under causal masking.

5 Attention Remasking

We introduce Attention Remasking (AR), a training-free edit to the attention score matrix that targets the multi-image failure induced by attention fragmentation. The objective is twofold: (i) reclaim attention currently assigned to visual sink tokens and (ii) counteract the biased attention distribution left by attention fragmentation, so that attention can flow along task-relevant cross-image links.

Motivation. Under causal masking, later images are inaccessible as keys to earlier queries, inherently restricting cross-image integration even when tasks explicitly require linking distributed visual evidence. To overcome this bottleneck, AR relaxes the visual components of $\mathcal{M}_{\text{causal}}$, permitting forward attention from earlier queries to subsequent visual tokens. Recent mechanistic studies reveal that LMMs inherently perform robust visual grounding in their earliest layers, prior to the severe representational degradation observed deeper in the network [37,38]. We posit that this depth-induced semantic drift is fundamentally driven by the attention fragmentation identified in our analysis: the progressive accumulation of attention sinks exacerbated by causal skew. Motivated by this insight, AR directly leverages the LMM's own shallow-layer attention pattern to guide the reassignment of sink attention to task-relevant visual tokens.

Masking attention sinks. Let Z^ℓ and α^ℓ be the pre-softmax scores and row-normalized attention weights defined in (1) and (2), respectively. Sink tokens are identified as in (3) and (4), yielding the per-layer set $\mathcal{S}^\ell \subset \{1, \ldots, N\}$. AR edits only the visual-visual submatrix of Z^ℓ, preserving text autoregression enforced by the causal mask $\mathcal{M}_{\text{causal}}$. For all layers $\ell > 1$, AR masks sink tokens by setting

$\tilde{Z}^{\ell}_{i,j} = -\infty$ for all i and all $j \in \mathcal{S}^{\ell}$, which implements a column-wise block on sinks and removes their incoming attention everywhere.

Unmasking visual tokens. To establish forward links to task-relevant patches, we extract the text-to-vision attention weights at the first decoder layer ($\ell = 1$) to obtain a semantic grounding prior uncorrupted by deep-layer fragmentation. For the set of text instruction tokens $\mathcal{T}$ and visual tokens $\mathcal{V}$, we define the global relevance of each non-sink visual token j as $P_{\text{global}}(j) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \alpha^{1}_{t,j}$. To unmask a sparse set of cross-image links, we employ mean attention thresholding, a well-established technique in dynamic token sparsification [8, 25]. To prevent massive sink activations from skewing the threshold, we define the valid visual candidate pool as $\mathcal{V}_{\text{valid}} = \mathcal{V} \setminus \mathcal{S}^{1}$. A token is deemed task-relevant if its global relevance exceeds the mean expectation of this valid pool, defined as $\mu_{\text{valid}} = \frac{1}{|\mathcal{V}_{\text{valid}}|}\sum_{k\in\mathcal{V}_{\text{valid}}} P_{\text{global}}(k)$. This dynamically isolates our unmasked candidate set $\mathcal{U} = \{j \in \mathcal{V}_{\text{valid}} \mid P_{\text{global}}(j) > \mu_{\text{valid}}\}$. For any visual query token i at deeper layers $\ell > 1$, let $\mathcal{U}_i$ denote the subset of tokens in $\mathcal{U}$ that belong to images subsequent to the query's own image. We normalize the global relevance scores over this subset to obtain a valid routing distribution $\pi_{i,j} = P_{\text{global}}(j)/\sum_{k\in\mathcal{U}_i} P_{\text{global}}(k)$ for $j \in \mathcal{U}_i$, with $\pi_{i,j} = 0$ otherwise. This yields a sparse, instruction-conditioned prior that natively concentrates on highly salient patches.

Attention redistribution. We reassign the attention released from sinks, guided by π. For query i, let the trapped sink attention at layer ℓ be $\eta^{\ell}_i = \sum_{j\in\mathcal{S}^{\ell}} \alpha^{\ell}_{i,j}$. After masking sinks, a plain softmax would implicitly redistribute this budget among the remaining keys. Instead, AR explicitly routes the same share to the newly opened, relevant links using π. Define the target row distribution $\hat{\alpha}^{\ell}_{i,\cdot}$ by:

$$\hat{\alpha}^{\ell}_{i,j} = (1-\eta^{\ell}_i)\,\frac{\alpha^{\ell}_{i,j}}{\sum_{k\notin\mathcal{S}^{\ell}\cup\mathcal{U}_i} \alpha^{\ell}_{i,k}}\,\mathbf{1}\big[j \notin \mathcal{S}^{\ell}\cup\mathcal{U}_i\big] \;+\; \eta^{\ell}_i\,\pi_{i,j}\,\mathbf{1}[j \in \mathcal{U}_i]\,. \tag{9}$$

We realize $\hat{\alpha}^{\ell}_{i,\cdot}$ by writing scores $\tilde{Z}^{\ell}_{i,j} = \log \hat{\alpha}^{\ell}_{i,j} + c^{\ell}_i$ for all non-$-\infty$ entries, where c^{ℓ}_i is any row-constant canceled by the softmax, while enforcing $-\infty$ on sink columns and on visual links that remain masked. If $\mathcal{S}^{\ell} = \varnothing$ or $\mathcal{U}_i = \varnothing$, AR reduces to the identity on row i.

6 Experiments

We evaluate Attention Remasking (AR) on interleaved-token LMMs to test whether it improves multi-image accuracy, reduces attention fragmentation, including lower late-layer entropy and reduced sensitivity to image order, and remains robust under ablations and controls.

6.1 Experimental Settings

Models. We evaluate a diverse range of open-source LMMs, including *LLaVA-OneVision-7B* [2], *Qwen3-VL-8B* [4], *InternVL3.5-8B* [33], and *DeepSeek-VL2-Small* [36]. All experiments are conducted on eight NVIDIA A100 GPUs.

Table 1: Average accuracy (%) on five multi-image benchmarks.

Method	Benchmark (Acc %)				
	MMIU	MuirBench	MIRB	LIBench	MIBench
Commercial API					
Gemini-3-Pro	68.1	65.7	72.5	70.3	69.9
GPT-5	65.3	65.9	69.6	70.2	70.0
Open Source					
LLaVA-OneVision-7B	36.9	35.0	47.9	53.5	53.8
+ VAR	38.0	35.5	48.7	54.2	55.0
+ SoFA	38.3	36.7	48.3	54.5	54.1
+ Ours	41.2	39.6	51.9	57.7	58.6
Qwen3-VL-8B	55.6	51.3	65.6	61.2	63.5
+ VAR	55.9	51.4	66.0	61.7	63.5
+ SoFA	56.7	52.3	66.2	61.6	64.8
+ Ours	59.4	54.9	69.2	65.5	67.6
InternVL3.5-8B	49.3	39.7	53.6	57.6	51.5
+ VAR	50.6	40.9	54.6	59.1	52.4
+ SoFA	50.5	41.3	54.9	59.3	52.7
+ Ours	54.1	43.8	57.9	61.9	55.6
DeepSeek-VL2-Small	51.4	40.5	45.8	58.3	53.7
+ VAR	52.1	41.1	46.6	59.4	54.6
+ SoFA	52.3	41.1	46.5	60.1	55.0
+ Ours	56.1	44.7	50.7	62.5	58.2

Baselines. We compare against two training-free post-hoc methods related to attention fragmentation. SoFt Attention (*SoFA*) linearly interpolates the attention between the standard causal attention and the bidirectional attention without masking, opening links from earlier queries to later images [28]. Visual Attention Redistribution (*VAR*) moves post-softmax attention mass from identified sink tokens to visual non-sink tokens [10].

Tasks and benchmarks. We adopt a comprehensive multi-image evaluation covering a wide range of complementary tasks. *MMIU* [16] covers seven relationship types with tasks that test comparison, retrieval, and spatial/temporal integration; *MuirBench* [30] spans twelve tasks built around relation categories such as multiview, ordering, and temporal reasoning; *MIRB* [40] groups tasks into perception, visual world knowledge, reasoning, and multi-hop reasoning; we also use the multi-image subset of *MIBench* [14], which focuses on five tasks of reasoning and comparison, and subset of LLaVA-Interleave Bench [12], spanning nine tasks on multi-image reasoning, which we refer to as *LIBench*.

6.2 Main Results

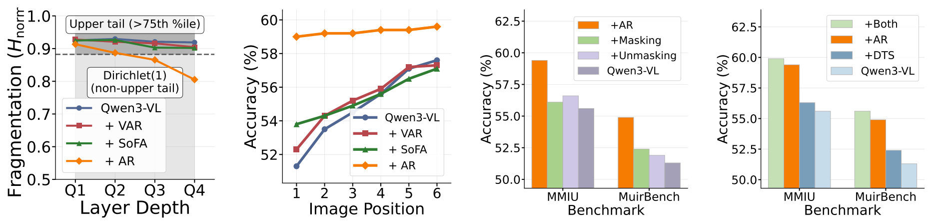
Overall accuracy. As shown in Tab. 1, AR improves average accuracy across multi-image benchmarks and open-source LMMs.

Level of fragmentation. Fig. 4a plots the normalized entropy H_{norm} by layer-depth quartiles. Baselines lie in the upper tail of a Dirichlet(1) compositional null across depth, indicating dispersed allocation over images. AR shifts the curve downward at all depths, with a statistically significant reduction in the final quartile as confirmed by paired Wilcoxon signed-rank tests with Holm correction. Fig. 4b shows accuracy as a function of image position under controlled permutations. Baselines display a clear positive slope, evidencing recency bias. AR both raises the curve and flattens the slope; permutation flip rates drop, with Wilson intervals remaining strictly below the baseline. For comparison baselines, SoFA reduces the position–accuracy slope while late-layer entropy remains high, and VAR leaves entropy unchanged and largely preserves the position–accuracy slope, reflecting that proportional post-softmax reweighting preserves the pre-existing fragmented pattern.

Discussion. SoFA [28] interpolates toward unmasked attention using logits that were never trained under the causal mask, opening all links without preference for task-relevant ones; this weakly counters positional bias but fails to consolidate attention, leaving fragmentation high. VAR [10] moves probability mass only within the already-masked distribution and in proportion to current non-sink weights; it cannot create forward links and therefore inherits both dispersion and positional skew. AR differs by masking sink keys, unmasking a sparse set of instruction-relevant cross-image links, and explicitly routing attention to those links, jointly reducing fragmentation and order sensitivity. Notably, while recent base LLMs such as Qwen3 have introduced architectural mitigations to prevent textual attention sinks with Gated Attention [24], we found that visual sinks still reliably emerge in their multimodal counterparts Qwen3-VL [4]. This persistence is driven by three intersecting factors: first, Vision Transformers inherently repurpose background patches as semantically empty registers [5]. Second, unlike text tokens, images largely consist of redundant backgrounds; to sharply focus on a few sparse foreground details, the softmax sum-to-one constraint forces the model to allocate residual probability mass to these empty patches. Finally, because the autoregressive loss is computed exclusively on text, visual tokens lack direct next-token gradient supervision, turning these unpenalized background registers into safe, low-penalty attention sinks. Because this trapped visual budget remains substantial, AR's mechanism of explicitly routing it through the causal mask to natively grounded forward links remains effective, delivering performance gains even on sink-aware baselines such as Qwen3-VL.

6.3 Further Analysis

Masking and unmasking. We compare two ablations against the full AR. *Mask-only* removes all incoming attention to identified sinks by setting their columns to $-\infty$ and then renormalizing each row over the remaining keys. This diffuses the budget that was trapped in sinks across the pre-existing non-sink pattern, so image-level dispersion and order sensitivity largely persist. *Unmask-only* relaxes the visual mask to open cross-image links and assigns scores equivalent to scores in sinks to those links in proportion to the relevance score, without



(a) (b) (c) (d)

Fig. 4: Attention Remasking (AR): results and ablations. (a) *Fragmentation:* AR lowers normalized entropy compared with baselines, indicating reduced attention fragmentation. (b) *Order sensitivity:* AR reduces position-wise accuracy skew relative to baselines. (c) *Ablation on masking strategy:* mask-only and unmask-only variants underperform full AR. (d) *Information leakage:* combining AR and Delimiter Token Scaling (DTS) maximizes accuracy, confirming orthogonal benefits.

Table 2: Ablation on the source layer used for the grounding prior. Extracting the Attention Remasking (AR) routing mask from the first layer maximizes accuracy. Conversely, using deep layers actively degrades performance below the baseline, empirically confirming that deep-layer attention is heavily corrupted by visual sinks.

Method	Source Layer	Benchmark (Acc %)				
		MMIU	MuirBench	MIRB	LIBench	MIBench
Qwen3-VL-8B	–	55.6	51.3	65.6	61.2	63.5
+ AR (Last)	L	53.2	49.8	63.1	59.5	61.0
+ AR (Mid)	$L/2$	54.8	50.9	64.8	60.6	62.4
+ AR (Ours)	1	**59.4**	**54.9**	**69.2**	**65.5**	**67.6**
InternVL3.5-8B	–	49.3	39.7	53.6	57.6	51.5
+ AR (Last)	L	46.5	37.4	51.0	55.2	48.9
+ AR (Mid)	$L/2$	48.7	39.1	52.8	56.9	50.8
+ AR (Ours)	1	**54.1**	**43.8**	**57.9**	**61.9**	**55.6**

masking and removing attention to sink tokens. Fig. 4c shows that for both variants, accuracy remains close to the baseline. The full AR masks sink and reallocates the freed budget to the newly unmasked, instruction-relevant links, and achieves the largest accuracy improvements among the variants.

Information leakage among images. In LMMs, processing multiple images within a single sequence can cause unintentional information leakage across images. To address this, Delimiter Token Scaling (DTS) reinforces intra-image boundaries by scaling the hidden states of image delimiter tokens [11]. Because DTS solely modifies these boundary representations while Attention Remasking (AR) edits the visual patch attention matrix, the two methods are structurally orthogonal. To evaluate their compatibility, we apply both interventions to the Qwen3-VL baseline. As shown in Fig. 4d, both methods individually improve multi-image reasoning accuracy on MMIU and MuirBench, with AR providing

Table 3: Average accuracy (%) on five multi-image benchmarks for larger models.

Method	Benchmark (Acc %)				
	MMIU	MuirBench	MIRB	LIBench	MIBench
Qwen3-VL-32B	59.8	54.8	68.7	66.0	66.7
+ VAR	60.1	55.0	69.1	66.3	67.0
+ SoFA	60.5	55.4	69.4	66.5	67.2
+ Ours	**62.6**	**57.5**	**71.8**	**68.8**	**69.4**
InternVL3.5-14B	53.0	44.6	57.0	61.7	56.1
+ VAR	53.4	45.1	57.5	62.0	56.6
+ SoFA	53.6	45.5	57.8	62.3	56.9
+ Ours	**56.1**	**47.9**	**60.2**	**64.7**	**59.3**
DeepSeek-VL2-27B	55.9	43.7	50.6	61.9	58.0
+ VAR	56.4	44.2	51.3	62.5	58.6
+ SoFA	56.7	44.6	51.5	62.8	58.9
+ Ours	**58.8**	**46.8**	**53.7**	**65.1**	**61.1**

notably larger standalone gains than DTS. Crucially, combining the two methods yields compounding improvements, achieving the highest overall accuracy across both benchmarks. This result confirms that preventing general noise leakage (DTS) and actively routing sink attention to task-relevant forward links (AR) provide independent, synergistic benefits, allowing AR to seamlessly integrate with state-of-the-art representation scaling techniques.

Computational Efficiency. We report single-stream latency (mean/p90 per sample) and offline throughput (samples/s) using CUDA-event timing after 20 warm-up runs and 200 measured runs. Hardware, precision, batch size, prompts, and images are strictly identical across methods. Peak memory is recorded via `torch.cuda.max_memory_allocated()`. The timing is conducted using the InternVL3.5-8B backbone on an NVIDIA A100 80 GB GPU (FP16). SoFA introduces the smallest overhead because it relies on a uniform, static interpolation of the attention mask. VAR redistributes mass dynamically based on existing non-sink weights, incurring a slight latency increase. The computational overhead of our proposed AR is limited to a lightweight extraction of the text-to-image attention weights at the first decoder layer, followed by a sparse, pre-softmax masking operation in deeper layers. A common concern with explicit weight extraction and custom masking is the potential incompatibility with hardware-fused kernels like *FlashAttention*, which typically do not materialize the full $N \times N$ matrix. To ensure efficiency, our implementation leverages PyTorch's `scaled_dot_product_attention` (SDPA). We strictly isolate the materialization of attention probabilities to the first layer. For all deeper layers, SDPA automatically routes the custom sparse AR mask to optimal Memory-Efficient Attention backends, bypassing the strict mask constraints of pure FlashAttention without incurring $O(N^2)$ memory bottlenecks. Consequently, AR operates with a latency, throughput, and memory footprint comparable to VAR and SoFA.

Table 4: Computational overhead on an NVIDIA A100. End-to-end latency includes image preprocessing, model forward, and decoding. AR introduces minimal overhead comparable to baselines.

Setting	Latency (ms) mean / p90	Throughput (samples/s)	Peak Mem (GB)
InternVL3.5-8B	**280 / 330**	**3.6**	**15.2**
+ SoFA	284 / 334 (+**1.4%**)	3.5 (–**2.8%**)	15.2 (+**0.0%**)
+ VAR	287 / 338 (+**2.5%**)	3.5 (–**2.8%**)	15.3 (+**0.7%**)
+ AR	288 / 339 (+**2.9%**)	3.5 (–**2.8%**)	15.3 (+**0.7%**)

Layer selection for the grounding prior. Attention Remasking (AR) leverages the earliest transformer layers, which preserve an uncorrupted, text-to-vision alignment prior [37, 38], whereas deep layers suffer from severe attention fragmentation. To validate this premise, we ablate the source layer used to extract the routing mask. As shown in Table 2, extracting the prior from the first layer (Layer 1) yields the highest performance gains across all benchmarks. However, as the source layer shifts deeper into the network (Layer $L/2$), the improvements vanish. Crucially, when the mask is extracted from the final layer (Layer L), the performance strictly degrades, falling noticeably below the unedited baseline. This behavior perfectly aligns with our mechanistic analysis of visual sinks: because deep-layer attention mass is largely absorbed by semantically empty background registers, thresholding this corrupted distribution generates an inverted mask that actively forces the model to focus on noise while blinding it to task-relevant foreground patches. Therefore, isolating the grounding prior to the first layer is not merely an optimization, but a structural requirement to bypass sink corruption and accurately route multi-image reasoning.

Scaling to larger backbones. Table 3 demonstrates the scalability of our method on larger-parameter models. While upgrading the base architectures to Qwen3-VL-32B [4], InternVL3.5-14B [33], and DeepSeek-VL2-27B [36] naturally yields higher baseline accuracy compared to their smaller counterparts, the underlying structural issue of visual attention fragmentation persists. Consequently, naive baseline edits such as VAR and SoFA struggle to meaningfully adjust the complex deep-layer representations, offering marginal gains. In contrast, Attention Remasking (Ours) consistently delivers robust improvements across all five multi-image benchmarks, regardless of the backbone.

7 Conclusion

We demonstrate that position-skewed visual sinks severely fragment attention in multi-image LMMs. To address this, we introduce Attention Remasking (AR), a training-free inference edit that suppresses these sinks and explicitly routes attention to task-relevant cross-image links. Future work will explore extending this framework to complex real-world applications of multi-image reasoning, such as industrial product specification, medical diagnostics, and animal re-identification [13, 18, 19, 21, 35].

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4190–4197. Association for Computational Linguistics, Online (Jul 2020). `https://doi.org/10.18653/v1/2020.acl-main.385`, `https://aclanthology.org/2020.acl-main.385/`
2. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
3. Araabi, A., Niculae, V., Monz, C.: Entropy–and distance-regularized attention improves low-resource neural machine translation. In: Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track). pp. 140–153 (2024)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
5. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
6. Di Zhao, J.Z., Hu, H., Fournier-Viger, P., Dobbie, G., Koh, Y.S.: Balancing invariant and specific knowledge for domain generalization with online knowledge distillation. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 2440–2448 (2025)
7. Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., Lin, M.: When attention sink emerges in language models: An empirical view. In: The Thirteenth International Conference on Learning Representations (2025), `https://openreview.net/forum?id=78Nn4QJTEN`
8. Hyeon-Woo, N., Yu-Ji, K., Heo, B., Han, D., Oh, S.J., Oh, T.H.: Scratching visual transformer's back with uniform attention. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5807–5818 (2023)
9. Jiang, D., He, X., Zeng, H., Wei, C., Ku, M.W., Liu, Q., Chen, W.: Mantis: Interleaved multi-image instruction tuning. Transactions on Machine Learning Research **2024** (2024), `https://openreview.net/forum?id=skLtdUVaJa`
10. Kang, S., Kim, J., Kim, J., Hwang, S.J.: See what you are told: Visual attention sink in large multimodal models. In: The Thirteenth International Conference on Learning Representations (2025), `https://openreview.net/forum?id=7uDI7w5RQA`
11. Lee, M., Park, Y., Hwang, D., Kim, Y., Oh, S.J., Choe, J.: Enhancing multi-image understanding through delimiter token scaling. arXiv preprint arXiv:2602.01984 (2026)
12. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., MA, Z., Li, C.: LLaVA-neXT-interleave: Tackling multi-image, video, and 3d in large multimodal models. In: The Thirteenth International Conference on Learning Representations (2025), `https://openreview.net/forum?id=oSQiao9GqB`
13. Li, Y., Zhao, D., Qiao, T., Wu, Y., Pang, B., Koh, Y.S.: Metawild: A multimodal dataset for animal re-identification with environmental metadata. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 13009–13015 (2025)
14. Liu, H., Zhang, X., Xu, H., Shi, Y., Jiang, C., Yan, M., Zhang, J., Huang, F., Yuan, C., Li, B., et al.: Mibench: Evaluating multimodal large language models over multiple images. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 22417–22428 (2024)

15. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024)
16. Meng, F., Wang, J., Li, C., Lu, Q., Tian, H., Yang, T., Liao, J., Zhu, X., Dai, J., Qiao, Y., Luo, P., Zhang, K., Shao, W.: MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=WsgEWL8i0K
17. Pang, B., Qiao, T., Walker, C., Cunningham, C., Koh, Y.S.: Cabin: Debiasing vision-language models using backdoor adjustments. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 484–492 (2025)
18. Pang, B., Qiao, T., Walker, C., Cunningham, C., Koh, Y.S.: Libra: Measuring bias of large language model from a local context. In: European Conference on Information Retrieval. pp. 1–16. Springer (2025)
19. Qi, H., Cao, J., Zhang, Y., Wang, X., Tang, W., Chen, B., Huo, C., Pan, H., You, H., Li, J., et al.: Industrybench-mipu: Benchmarking multi-image attribute value extraction for industrial products. arXiv preprint arXiv:2606.14383 (2026)
20. Qiao, T., Walker, C., Cunningham, C., Jang-Jones, A., Morton, S., Meissel, K., Koh, Y.S.: Thematic bottleneck models for multimodal analysis of school attendance. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management. pp. 5971–5979 (2025)
21. Qiao, T., Walker, C., Cunningham, C., Jang-Jones, A., Morton, S., Meissel, K., Sing Koh, Y.: Longitudinal surveys are texts: Llm-enhanced analysis of school attendance in new zealand. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 310–327. Springer (2025)
22. Qiao, T., Walker, C., Cunningham, C., Koh, Y.S.: Thematic-lm: a llm-based multi-agent system for large-scale thematic analysis. In: Proceedings of the ACM on Web Conference 2025. pp. 649–658 (2025)
23. Qiao, T., Zhao, D., Walker, C., Cunningham, C., Koh, Y.S.: Zero-shot domain generalisation via prompt-driven feature refinement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6184–6193 (2026)
24. Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., et al.: Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. arXiv preprint arXiv:2505.06708 (2025)
25. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems **34**, 13937–13949 (2021)
26. Sun, M., Chen, X., Kolter, J.Z., Liu, Z.: Massive activations in large language models. In: First Conference on Language Modeling (2024), https://openreview.net/forum?id=F7aAhfitX6
27. Team, Q.: Qwen2.5 technical report (2025), https://arxiv.org/abs/2412.15115
28. Tian, X., Zou, S., Yang, Z., Zhang, J.: Identifying and mitigating position bias of multi-image vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10599–10609 (2025)
29. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5797–5808. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1580, https://aclanthology.org/P19-1580/

30. Wang, F., Fu, X., Huang, J.Y., Li, Z., Liu, Q., Liu, X., Ma, M.D., Xu, N., Zhou, W., Zhang, K., Yan, T.L., Mo, W.J., Liu, H.H., Lu, P., Li, C., Xiao, C., Chang, K.W., Roth, D., Zhang, S., Poon, H., Chen, M.: Muirbench: A comprehensive benchmark for robust multi-image understanding. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=TrVYEZtSQH
31. Wang, H., Wu, Z., Guo, J., Han, W., Liu, L., Yang, Q., Shao, J.: Exploiting reliable evolving micro-clusters for robust semi-supervised learning on data streams. Information Sciences p. 123069 (2026)
32. Wang, H., Wu, Z., Guo, J., Hao, Q., Liu, X., Liu, H., Yang, Q., Shao, J.: Learning contrastive evolving micro-clusters for robust semi-supervised data stream classification. IEEE Transactions on Cybernetics (2025)
33. 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)
34. Wu, X., Wang, Y., Jegelka, S., Jadbabaie, A.: On the emergence of position bias in transformers. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=YufVk7I6Ii
35. Wu, Y., Zhao, D., Li, Y., Alajas, M., Glen, A.S., Zhang, J., Dobbie, G., Wilson, D., Koh, Y.S.: Overcoming fine-grained visual challenges in animal re-identification via semantic feature alignment. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 371–381 (2026)
36. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
37. Yoon, H., Jung, J., Kim, J., Choi, H., Shin, H., Lim, S., An, H., Kim, C., Han, J., Kim, D., et al.: Visual representation alignment for multimodal large language models. arXiv preprint arXiv:2509.07979 (2025)
38. Yu, Z., Lee, Y.J.: How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. arXiv preprint arXiv:2508.20279 (2025)
39. Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., Susskind, J.M.: Stabilizing transformer training by preventing attention entropy collapse. In: International Conference on Machine Learning. pp. 40770–40803. PMLR (2023)
40. Zhao, B., Zong, Y., Zhang, L., Hospedales, T.: Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. arXiv preprint arXiv:2406.12742 (2024)
41. Zhao, D., Koh, Y.S., Dobbie, G., Hu, H., Fournier-Viger, P.: Symmetric self-paced learning for domain generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 16961–16969 (2024)
42. Zhao, D., Zhang, J., Hu, H., Fournier-Viger, P., Dobbie, G., Koh, Y.S.: Unlearning during training: Domain-specific gradient ascent for domain generalization. In: The Fourteenth International Conference on Learning Representations (2026)