AMR: ENHANCING VISUAL EVIDENCE SELECTION VIA ATTENTION-GUIDED MULTIMODAL RE-READING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have advanced rapidly across vision-language tasks, yet their performance on visually grounded reasoning remains far from perfect. We investigate whether repeating the same image–question pair can address a limitation of causal, left-to-right processing, whereby visual evidence may be encountered before the model fully understands which information is relevant to the question. Repeating the complete multimodal input consistently improves inference, and an analysis of the repeated pass reveals a systematic redistribution of text-to-vision attention between its first and second reads. We refer to this read-to-read redistribution as *attention shift*. Rather than uniformly replaying the first allocation, the second read selectively increases attention to a subset of visual tokens, indicating that repetition can expose complementary visual evidence. Building on this observation, we propose *AMR* (Attention-Guided Multimodal Re-Reading), a training-free, plug-and-play intervention that leverages the model's own attention shift to identify and selectively reinforce visual evidence that becomes more relevant during the second read. Across six MLLMs and four multimodal reasoning benchmarks, AMR improves all 24 model–benchmark combinations over the zero-shot baseline, with an average gain of 2.61 percentage points. These results suggest that multimodal re-reading provides more than additional computation: it yields a state-dependent signal that can guide visual evidence selection at inference time, without parameter updates or task-specific training.

1 INTRODUCTION

Multimodal large language models (MLLMs) have advanced on vision–language tasks, yet their performance on visually grounded reasoning remains far from perfect (Alayrac et al., 2022; Li et al., 2023; Dai et al., 2023; Liu et al., 2023; 2024a; Chen et al., 2024b). We investigate this gap through a limitation of causal multimodal processing. In decoder-based MLLMs, visual and textual tokens are processed under a causal attention mask in a single left-to-right order (Bai et al., 2025; Zhu et al., 2025; Team et al., 2025). Each token can access only preceding multimodal context, so visual evidence may be encountered before the model understands the question. Once the reasoning state changes, earlier evidence cannot be revisited. This fixed order is misaligned with complex problem solving: evidence whose relevance is initially unclear may become decisive only after an interpretation has formed. Consequently, visual evidence can be overlooked during the initial read.

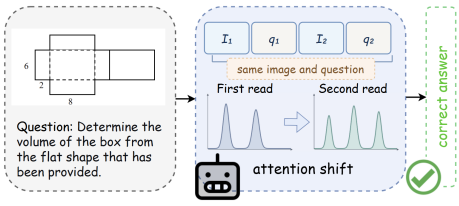

Figure 1: **Teaser: state-dependent visual re-reading.** The second read reallocates visual attention after the first read has updated the multimodal state, revealing complementary evidence for the answer.

Most prompting methods allocate additional computation to the output process, for example by eliciting intermediate reasoning steps, sampling multiple reasoning paths, or planning a solution before execution (Wei et al., 2022; Wang et al., 2023b;a). A complementary line of work shows that the organization of the input itself can also affect reasoning (Cheng et al., 2024). In particular, Re-Reading (RE2) improves reasoning in autoregressive language models by presenting the same input twice (Xu et al., 2024). One interpretation is that the first read establishes

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

an intermediate state, while the second can condition on information accumulated during that read. For MLLMs, this suggests that repeating the image–question pair may let the model revisit visual evidence after its understanding of the problem has been updated. Re-reading may therefore help not because evidence is processed twice, but because the second read occurs under a globally informed internal state.

We test this hypothesis in the multimodal setting and find that repeating the complete image–text input improves model performance. This result raises a question: what changes when visual evidence is encountered for a second time? As illustrated in Figure 1, the second read can alter the visual evidence used to solve the same problem, rather than merely replaying the first computation.

We therefore focus on the repeated multimodal pass and compare the visual attention distributions produced by its first and second reads. This read-to-read comparison reveals a token-level redistribution of normalized text-to-vision attention: some visual tokens receive more attention during the second read, whereas others do not exhibit the same increase. We call this measurable redistribution an *attention shift*. The second read is thus not a uniform replay of the first, but a re-evaluation of visual evidence after the model has formed a more complete interpretation.

Based on this observation, we introduce *AMR* (*Attention-Guided Multimodal Re-Reading*), a training-free, plug-and-play inference intervention. It tracks how visual attention changes between the two reads, identifies candidate evidence whose relevance increases, and selectively reinforces that evidence during decoding. Repeat therefore serves not only as an effective inference strategy, but also as a source of a model-internal signal that can be operationalized at test time. Our contributions are summarized as follows:

- We show that repeating a complete multimodal input can improve reasoning beyond a single read, indicating that a globally updated multimodal context can be useful for test-time inference.
- We discover and characterize *attention shift* in Repeat and propose *AMR*, a training-free, plug-and-play inference intervention that selectively reinforces visual evidence whose relevance increases across repeated reads.
- We evaluate the method across six MLLMs and four multimodal reasoning benchmarks, obtaining consistent gains across model scales, architectures, and task types.

2 PRELIMINARIES

2.1 CAUSAL ATTENTION IN DECODER-BASED MLLMS

An autoregressive language model generates each output token conditioned on the preceding sequence. For an output sequence $y_{1:m}$ and an input prefix X, its conditional distribution factorizes as

$$P(y_{1:m} \mid X) = \prod_{t=1}^{m} P(y_t \mid X, y_{<t}). \tag{1}$$

Within the language decoder, a causal attention mask likewise restricts each position to attending to itself and earlier positions. Writing M for this mask, $M_{ij} = 0$ when $j \leq i$ and $M_{ij} = -\infty$ otherwise. Thus, a representation formed at an earlier position cannot be updated using tokens that appear later in the sequence.

In the decoder-based MLLMs studied here, visual tokens and question tokens participate in this left-to-right attention process. Consequently, visual evidence encountered early in an image–question sequence cannot be recontextualized at that position using the question information processed later. This does not prevent the model from using the full input when generating an answer; it motivates a second encounter with the visual evidence after the first image–question interaction.

2.2 REPEATED MULTIMODAL READING

Let I be an image and q its question. Standard inference presents the pair once as $[I, q]$. We define REPEAT by presenting the same pair twice:

$$X_{\text{rep}} = [I_1, q_1, I_2, q_2], \qquad I_1 = I_2 = I, \quad q_1 = q_2 = q. \tag{2}$$

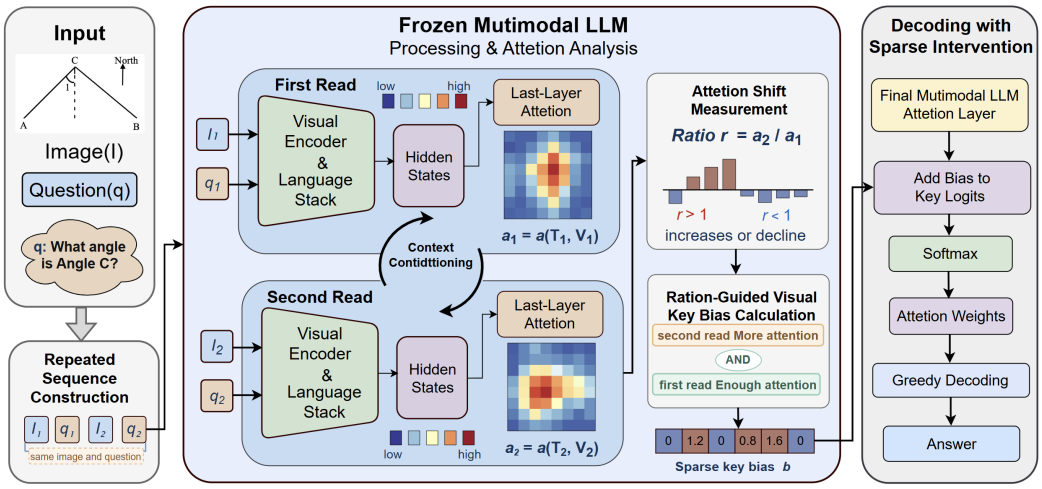

Figure 2: **Overview of *AMR*.** The same image–question pair is incorporated twice within a single input sequence to a frozen MLLM. The first read establishes an initial multimodal state, and the second revisits the image after that state has been updated. We compare normalized text-to-vision attention across the two reads, select visual tokens whose supported attention increases, and convert their attention ratios into a sparse key-position bias. The bias is applied to the second visual block during greedy decoding.

We refer to $[I_1, q_1]$ and $[I_2, q_2]$ as the first and second reads, respectively. Although the image and question are unchanged, the second read follows the complete first read and therefore has access to a different causal context. Under identical preprocessing and model settings, the first read has the same causal prefix as the one-pass input $[I, q]$. This correspondence links first-read attention to the Original baseline, while the attention shift studied below is defined between the two reads within REPEAT.

3 METHOD

3.1 OVERVIEW AND FORMULATION

Given an image–question pair (I, q) and a frozen MLLM f_θ, our goal is to use the change in visual attention across the two reads as an inference-time signal. We seek an answer

$$\hat{y} = g_\alpha(f_\theta; I, q), \tag{3}$$

where g_α denotes the proposed inference procedure and α controls the strength of its attention intervention. The model parameters θ remain fixed.

Figure 2 summarizes *AMR* in four stages. *Repeat* provides two reads of the same multimodal input. *Compare* measures the change in normalized text-to-vision attention between them. *Select and Enhance* converts supported attention increases into a sparse visual-token signal. Finally, *Inject* applies this signal as a key-position bias to the second visual block during autoregressive decoding. The procedure requires no additional training and leaves the visual encoder, tokenizer, language-model weights, and output head unchanged.

3.2 STAGE 1: REPEATED READING FOR MULTIMODAL STATE REFRESH

The left side of Figure 2 shows the repeated reading stage. Under causal decoding, the first image–question block is processed without access to the second. By contrast, the second block can condition on the full prefix produced by the first image–question interaction. The two reads therefore share visual content and question semantics but differ in their available context. This controlled construction isolates how the model's allocation of attention to unchanged visual evidence evolves after one multimodal read.

3.3 Stage 2: Measuring Read-to-Read Attention Shift

We next measure how visual evidence is reweighted across the two reads, as shown in the center of Figure 2. Let $\mathcal{V}_1$ and $\mathcal{V}_2$ be the visual-token position sets associated with I_1 and I_2, and let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the text-token position sets associated with q_1 and q_2. For a text-query set $\mathcal{T}$ and a visual-key set $\mathcal{V}$, we define the normalized text-to-vision attention distribution

$$a(\mathcal{T},\mathcal{V})_j = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \frac{\frac{1}{H}\sum_{h=1}^{H} A_{h,t,j}}{\sum_{k\in\mathcal{V}}\frac{1}{H}\sum_{h=1}^{H} A_{h,t,k}}, \qquad j\in\mathcal{V}. \tag{4}$$

where $A_{h,t,j}$ is the final-layer attention probability from text query t to visual key j in head h, and H is the number of attention heads. This quantity first averages across heads, normalizes the attention mass within the visual block for each text query, and then averages across text queries. The two read-specific distributions are

$$a_1 = a(\mathcal{T}_1,\mathcal{V}_1), \qquad a_2 = a(\mathcal{T}_2,\mathcal{V}_2). \tag{5}$$

Because I_1 and I_2 are identical, their visual tokens are aligned by relative position. For each aligned position j, we define the attention-shift ratio

$$r_j = \frac{a_{2,j}+\epsilon}{a_{1,j}+\epsilon}, \tag{6}$$

where $\epsilon > 0$ is a fixed numerical-stability constant. A value $r_j > 1$ means that the corresponding visual position receives greater normalized attention during the second read.

3.4 Stage 3: Ratio-Guided Visual Evidence Selection and Enhancement

Large ratios can arise from near-zero first-read attention and need not provide a reliable enhancement signal. As shown on the right side of Figure 2, we therefore retain only attention increases supported by sufficient first-read mass. We define

$$\tau = \mathrm{median}_j(a_{1,j}), \tag{7}$$

and define the selected visual-token set as

$$\mathcal{S} = \{j \mid r_j > 1 \ \wedge \ a_{1,j} \geq \tau\}. \tag{8}$$

The ratio condition identifies evidence that becomes more prominent on the second read, while the median threshold suppresses unstable changes from weakly supported tokens. We assign each selected position an additive enhancement

$$b_j = \begin{cases} \alpha \log r_j, & j \in \mathcal{S}, \\ 0, & j \notin \mathcal{S}. \end{cases} \tag{9}$$

Here, $\alpha \geq 0$ controls the intervention strength. The logarithm maps a multiplicative attention ratio to an additive quantity in logit space and compresses unusually large ratios. The result is a sparse bias vector: tokens outside $\mathcal{S}$ are left unchanged.

3.5 Stage 4: Injecting Visual Key Bias for Autoregressive Decoding

Finally, the right side of Figure 2 maps b to the key positions of the second visual block $\mathcal{V}_2$. Let $z_{h,t,j}$ denote the original attention logit in head h from query t to key j. We modify only keys in the second visual block:

$$\widetilde{z}_{h,t,j} = z_{h,t,j} + b_j \, \mathbf{1}[j \in \mathcal{V}_2]. \tag{10}$$

The bias is broadcast across eligible queries and attention heads after causal masking and before softmax. Selected keys receive a positive offset, while all other positions retain their original logits. The frozen model then performs greedy autoregressive decoding under the intervened attention distribution to produce the final answer. *AMR* therefore uses the model's own read-to-read attention change without introducing new parameters or training objectives.

Table 1: Performance (%) across six MLLMs and four multimodal reasoning benchmarks. We report TreeBench accuracy, WeMath Strict and CM-Strict accuracy, CharXiv Reasoning accuracy, and MMBench circular accuracy. Each block presents the zero-shot baseline, *AMR*, and the absolute gain (Δ). *Mean* averages the four primary scores and excludes the auxiliary WeMath CM-Strict metric, while *Macro average* denotes the average across all six models. Red arrows denote improvements over the baseline.

Model	Params.	Setting	TreeBench	WeMath		CharXiv-R	MMBench	Mean
			Acc.	Strict	CM-Strict	Acc.	Acc.	
Qwen2.5-VL	7B	Baseline	27.41	36.86	29.71	41.20	75.69	45.29
		Ours	**33.09**	**39.71**	**32.57**	**43.40**	**79.38**	**48.90**
		Δ	**+5.68** ↑	**+2.85** ↑	**+2.86** ↑	**+2.20** ↑	**+3.69** ↑	**+3.61** ↑
	32B	Baseline	34.32	50.67	44.76	41.20	82.13	52.08
		Ours	**35.80**	**52.00**	**46.29**	**46.40**	**85.48**	**54.92**
		Δ	**+1.48** ↑	**+1.33** ↑	**+1.53** ↑	**+5.20** ↑	**+3.35** ↑	**+2.84** ↑
InternVL3	8B	Baseline	38.77	34.95	27.43	36.20	83.25	48.29
		Ours	**40.25**	**38.00**	**30.48**	**37.90**	**85.48**	**50.41**
		Δ	**+1.48** ↑	**+3.05** ↑	**+3.05** ↑	**+1.70** ↑	**+2.23** ↑	**+2.12** ↑
	38B	Baseline	37.78	42.29	34.48	43.10	89.52	53.17
		Ours	**38.02**	**44.10**	**37.33**	**49.10**	**90.46**	**55.42**
		Δ	**+0.24** ↑	**+1.81** ↑	**+2.85** ↑	**+6.00** ↑	**+0.94** ↑	**+2.25** ↑
Gemma3	4B	Baseline	24.69	23.43	17.33	20.50	66.41	33.76
		Ours	**28.40**	**24.48**	**17.71**	**22.50**	**69.59**	**36.24**
		Δ	**+3.71** ↑	**+1.05** ↑	**+0.38** ↑	**+2.00** ↑	**+3.18** ↑	**+2.48** ↑
	27B	Baseline	30.37	41.24	35.81	31.40	80.58	45.90
		Ours	**32.59**	**44.29**	**38.86**	**34.20**	**82.04**	**48.28**
		Δ	**+2.22** ↑	**+3.05** ↑	**+3.05** ↑	**+2.80** ↑	**+1.46** ↑	**+2.38** ↑
Macro average	all models	Baseline	32.22	38.24	31.59	35.60	79.60	46.42
		Ours	**34.69**	**40.43**	**33.87**	**38.92**	**82.07**	**49.03**
		Δ	**+2.47** ↑	**+2.19** ↑	**+2.29** ↑	**+3.32** ↑	**+2.48** ↑	**+2.61** ↑

4 EXPERIMENTS

In this section, we systematically evaluate the effectiveness and underlying mechanism of *AMR* for multimodal reasoning. Specifically, we address three research questions: **(RQ1)** Can *AMR* consistently improve multimodal reasoning across MLLM architectures, model scales and task types? **(RQ2)** What roles do multimodal repetition and attention-guided reinforcement play in the observed gains? **(RQ3)** How does *AMR* change the visual evidence selected during inference? We first introduce the experimental setup and then answer these questions through quantitative comparisons, controlled ablations and visual attention analysis.

4.1 EXPERIMENTAL SETUP

4.1.1 MODELS

We evaluate our method on six multimodal large language models (MLLMs) from three model families: Qwen2.5-VL-7B and Qwen2.5-VL-32B (Bai et al., 2025), InternVL3-8B and InternVL3-38B (Zhu et al., 2025), and Gemma3-4B and Gemma3-27B (Team et al., 2025). This collection spans three widely used open MLLM architectures and model scales from 4B to 38B parameters, allowing us to test whether the proposed inference intervention transfers across model families and capacities.

Table 2: **Ablation of visual and textual repetition on Qwen2.5-VL-7B (%).** We compare ORI, Repeat Text, Repeat Image, REPEAT, and Full. Repeat Text and Repeat Image duplicate only the text or image block, respectively; REPEAT duplicates the full image–text input without attention-guided reinforcement, while Full further applies ratio-guided reinforcement to the second visual block. Best results are bold.

Setting	TreeBench	WeMath		CharXiv-R	MMBench
	Acc.	Strict	CM-Strict	Acc.	Acc.
ORI	27.41	36.86	29.71	41.20	75.69
Repeat Text	31.11	36.57	28.95	42.70	78.09
Repeat Image	31.36	36.38	28.95	41.10	72.68
REPEAT	29.38	37.90	30.86	42.50	79.30
Full	**33.09**	**39.71**	**32.57**	**43.40**	**79.38**

4.1.2 DATASETS AND TASKS

We evaluate four complementary multimodal benchmarks spanning grounded reasoning, mathematical reasoning, chart understanding, and general vision–language capabilities. TreeBench (Wang et al., 2026) focuses on visually grounded reasoning and reports accuracy. We-Math (Qiao et al., 2025) evaluates multimodal mathematical reasoning with Strict and CM-Strict accuracy. CharXiv (Wang et al., 2024) assesses chart understanding on its more challenging Reasoning subset, for which we report accuracy. MMBench (Liu et al., 2024b) evaluates general vision–language capabilities with circular accuracy. All results follow the official evaluation protocols and are reported as percentages.

4.2 RQ1: CAN *AMR* IMPROVE MULTIMODAL REASONING?

To evaluate whether *AMR* improves multimodal reasoning across architectures and scales, we compare the full method with the standard one-pass zero-shot baseline. Table 1 reports all 24 model–benchmark comparisons, together with the complementary WeMath CM-Strict metric. Means use WeMath Strict to avoid counting the same dataset twice.

The method improves all 24 primary comparisons, by 2.61 percentage points on average. Mean gains are 2.47 on TreeBench, 2.19 on WeMath Strict, 3.32 on CharXiv-Reasoning, and 2.48 on MMBench; CM-Strict also improves for every model (2.29 points on average). The all-positive pattern spans architectures and task types, but the variation in gain matters: the intervention has a larger average effect on chart reasoning than on mathematical reasoning.

Within Qwen2.5-VL, the 7B model gains 5.68 points on TreeBench (27.41% to 33.09%) and 3.69 on MMBench (75.69% to 79.38%), while also improving WeMath Strict and CharXiv-Reasoning by 2.85 and 2.20 points. The 32B model retains positive gains on all tasks, most notably 5.20 points on CharXiv-Reasoning (41.20% to 46.40%). Thus, benefits persist at the larger scale rather than being restricted to the 7B model.

The same cross-scale pattern holds for InternVL3 and Gemma3. InternVL3-8B gains 3.05 points on WeMath Strict and 2.23 on MMBench; InternVL3-38B gains 6.00 on CharXiv-Reasoning (43.10% to 49.10%), the largest single gain in Table 1, but only 0.24 on TreeBench. Gemma3-4B improves by 3.71 points on TreeBench and 3.18 on MMBench, whereas Gemma3-27B improves by 3.05 on WeMath Strict and 2.80 on CharXiv-Reasoning. Gains therefore extend across all three families and both evaluated scales in each family, without implying that their magnitudes are uniform.

The task pattern is compatible with the method's motivation. Charts and grounded reasoning can require the model to select and combine distinct visual regions after interpreting the question, so a state-conditioned second read offers another opportunity to allocate visual attention. This is an interpretation of the aggregate results, not evidence that every improved answer follows this mechanism; RQ2 isolates the intervention's components and RQ3 examines attention in individual cases.

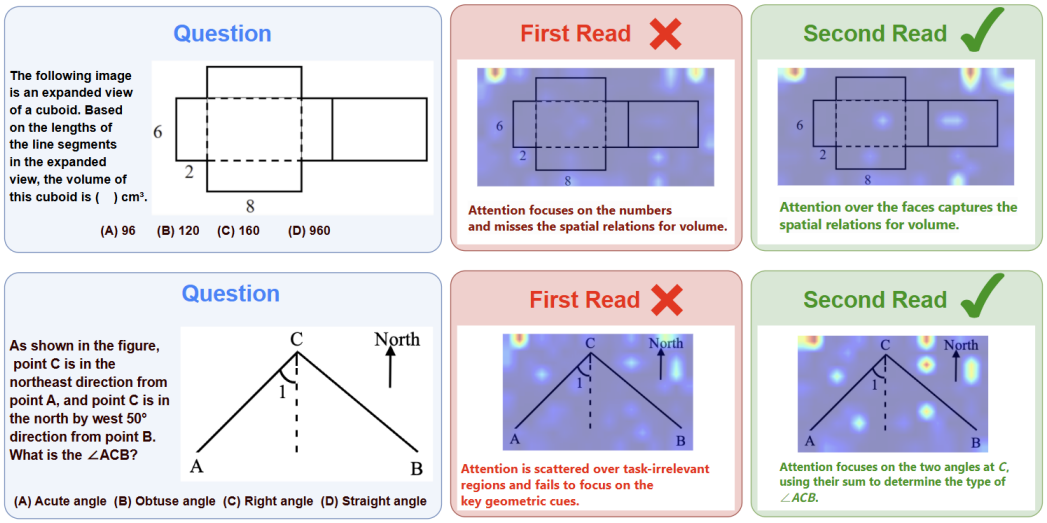

Figure 3: **Attention shifts between the first and second reads on two We-Math reasoning problems.** The top row shows a cuboid-net volume problem, where attention shifts from isolated numerical cues in the first read to the connected rectangular faces that define the spatial structure in the second read. The bottom row shows a directional-angle problem, where attention over task-relevant visual regions becomes more focused in the second read, concentrating on the two rays meeting at C and their associated angles. Across both examples, repetition redirects attention toward task-relevant visual evidence, revealing a consistent attention-shift pattern between the two reads.

4.3 RQ2: What Drives the Gains of *AMR*?

To answer RQ2, we conduct controlled ablations that progressively isolate the effects of modality repetition and attention-guided reinforcement (Table 2). We compare five settings: **Original**, the standard one-pass image–text baseline; **Repeat Text**, which duplicates the text while keeping one image block; **Repeat Image**, which duplicates the image while keeping one text block; **Repeat**, which duplicates the complete image–text input without attention-guided reinforcement; and **Full**, our complete *AMR* method. This ordering allows us to distinguish one-sided repetition from joint multimodal state refresh and to measure the additional contribution of the attention-guided intervention. The one-sided variants have task-dependent effects. Repeat Text increases TreeBench from 27.41% to 31.11% and CharXiv-Reasoning from 41.20% to 42.70%, but lowers WeMath Strict from 36.86% to 36.57%. Repeat Image raises TreeBench to 31.36%, yet falls below Original on WeMath (36.38% versus 36.86%), CharXiv-Reasoning (41.10% versus 41.20%), and especially MMBench (72.68% versus 75.69%). Hence neither repeating question tokens alone nor presenting the image twice reliably benefits all tasks. The MMBench drop is particularly important: increased visual exposure by itself can even hurt option selection.

Joint image–text repetition is more reliable. Without any attention-guided bias, Repeat improves Original by 1.97 points on TreeBench, 1.04 on WeMath Strict, 1.30 on CharXiv-Reasoning, and 3.61 on MMBench. Its 79.30% MMBench score contrasts sharply with the 72.68% of Repeat Image. The comparison is consistent with a second visual read benefiting from the accompanying question context, although it does not by itself isolate the internal cause of that benefit.

Adding ratio-guided reinforcement improves Repeat further on all four tasks: by 3.71 points on TreeBench (29.38% to 33.09%), 1.81 on WeMath Strict (37.90% to 39.71%), 0.90 on CharXiv-Reasoning (42.50% to 43.40%), and 0.08 on MMBench (79.30% to 79.38%). The TreeBench difference is consistent with selective visual reweighting being useful when several grounded cues must be combined; it also shows that the Full gain cannot be attributed to repeated input alone. WeMath CM-Strict follows the same progression: Original scores 29.71%, Repeat scores 30.86%, and Full scores 32.57%, whereas both one-sided variants score 28.95%.

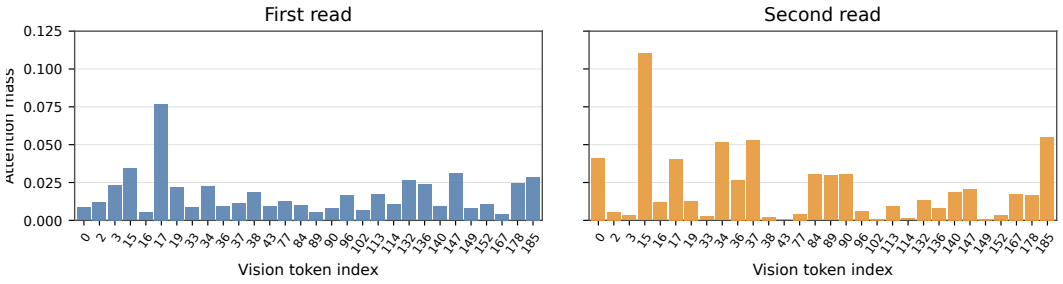

Figure 4: **Read-to-read attention redistribution in the cuboid-net problem.** We retain the 30 visual tokens with the largest absolute first-to-second-read attention differences and show them in the same token-index order in both panels. The dominant peak moves from token 17 to token 15, while tokens 0, 34, 37, and 185 gain attention. These changes reflect redistribution across the same visual tokens, not uniform rescaling.

The small Full–Repeat margin on MMBench qualifies this result. Its multiple-choice protocol requires an option rather than an explicit reasoning trace, and repetition alone accounts for nearly all of the observed gain. Task format is one plausible explanation, but these ablations cannot establish it as the cause. Overall, the sequence of comparisons separates the reliable effect of joint repetition from the additional, task-dependent contribution of attention-guided visual reinforcement.

4.4 RQ3: How Does *AMR* Change Visual Evidence Selection?

We analyze two representative WeMath problems to examine how visual evidence selection changes across repeated reads. One asks for the volume of a cuboid from its planar net; the other asks for the type of $\angle ACB$ from two directional constraints. Both require combining spatial relations rather than reading isolated symbols. For each problem, we compare attention within Repeat but relate the first-read attention map to the answer from the one-pass Original baseline. Under matched preprocessing and model settings, the first image–question block has the same causal context as Original and thus yields the same attention distribution. This correspondence lets us examine the baseline answer alongside first-read attention without treating that answer as an output of the first read itself.

Cuboid-net volume. In the first read (top-middle panel of Figure 3), attention is concentrated on individual numbers and short line segments, with limited coverage of the rectangular faces. The corresponding Original baseline treats $8{+}2 = 10$ as one dimension and predicts $V = 10{\times}6{\times}2 = 120$, the incorrect option B. Although it uses numbers shown in the image, this calculation does not recover their roles in the net: 8 and 2 belong to different parts of the planar layout and should not be combined into one dimension.

In the second read (top-right panel), attention is distributed across several faces of the net. This pattern is consistent with considering the net as a foldable three-dimensional object rather than combining nearby numerical labels. *AMR* assigns the dimensions $(8, 6, 2)$, computes $V = 8 \times 6 \times 2 = 96$, and selects the correct option A.

Directional-angle reasoning. The second example exhibits a related change in visual selection. First-read attention (bottom-middle panel) is dispersed across regions that do not consistently connect the two directional cues at C. The corresponding Original baseline predicts an acute angle (option A). On the second read (bottom-right panel), attention becomes more concentrated around the rays meeting at C and their directional cues. *AMR* uses the $45°$ northeast direction from A and the $50°$ north-by-west direction from B to identify $\angle ACB$ as obtuse, yielding the correct option B.

Quantifying the redistribution. Figure 4 visualizes the change for the cuboid-net example. Across its full visual-token distributions, first- and second-read attention have cosine similarity 0.789, Jensen–Shannon divergence 0.101, and ℓ_1 distance 0.734. The positive attention difference mass is 0.367,

meaning that approximately 37% of the normalized visual attention mass is redistributed toward tokens whose attention increases. Thus, the two reads remain related but do not use an identical allocation.

The change is selective. Only 21 of 190 visual tokens (11.1%) meet the method's selection rule, and their mean second-to-first attention ratio is 2.54. By comparison, the median ratio across all visual tokens is 0.56, and only 25.3% have a ratio above one. The top-10% attention sets have a Jaccard overlap of 0.36 between reads, further indicating that the most attended tokens change. Together, the case analysis and these measurements show that repeated reading yields a structured attention shift that *AMR* can use to selectively reinforce visual evidence.

5 Related Work

5.1 Input-Side Computation for Language-Model Reasoning

Most prompting methods enhance language-model reasoning by allocating additional computation during output generation, such as chain-of-thought prompting for intermediate reasoning, self-consistency for aggregating multiple trajectories, and plan-and-solve prompting for separating planning from execution (Wei et al., 2022; Wang et al., 2023b;a). A complementary line of work improves reasoning by modifying the input itself: Information Re-Organization restructures the given information, while Re-Reading (RE2) repeats the input so that the second occurrence can condition on information accumulated during the first (Cheng et al., 2024; Xu et al., 2024). We extend this input-side perspective to multimodal reasoning by analyzing how text-to-vision attention changes across repeated reads and using the resulting shift as an inference signal.

5.2 Multimodal Reasoning and Visual Evidence Selection

Modern MLLMs connect visual encoders with autoregressive language models through cross-attention, learned projections, or lightweight interfaces (Alayrac et al., 2022; Li et al., 2023; Dai et al., 2023; Liu et al., 2023; 2024a; Chen et al., 2024b). Existing methods structure multimodal reasoning through role separation, iterative decomposition, visual grounding, or visual prompting (Zheng et al., 2023; You et al., 2023; Chen et al., 2024c; Lu et al., 2025). Tool-based approaches instead acquire new crops, invoke external modules, or construct intermediate visual representations (Gupta & Kembhavi, 2023; Suris et al., 2023; Lu et al., 2023; Wu & Xie, 2024; Hu et al., 2024). *AMR* differs by keeping underlying image-question content and frozen MLLM unchanged while examining how visual attention evolves across repeated reads.

Visual-token efficiency methods estimate token importance within a single inference trajectory for pruning, retention, or compression (Cao et al., 2023; Chen et al., 2024a; Yang et al., 2025; Zhuang et al., 2025). In contrast, we measure normalized attention changes for corresponding tokens across two state-conditioned reads and selectively reinforce emerging evidence. Since attention is not necessarily a faithful causal explanation (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), we treat attention shift as an internal signal and validate it through an explicit decoding intervention.

6 Conclusion

We show that repeating the same image–question input enables an MLLM to revisit visual evidence under an updated multimodal state. Our analysis reveals a systematic attention shift between the first and second reads, in which the model reallocates attention toward visual evidence that becomes more relevant after the initial read. Building on this observation, we propose *AMR*, a training-free inference method that selectively reinforces such evidence without modifying model parameters. Extensive experiments across six MLLMs and four multimodal reasoning benchmarks demonstrate consistent improvements over standard one-pass inference, while controlled ablations confirm the complementary roles of multimodal repetition and attention-guided reinforcement. These findings establish repeated reading as an effective inference-time strategy and suggest that state-dependent attention changes provide a useful internal signal for improving visual evidence selection in multimodal reasoning.

Reproducibility statement

Our method is inference-only and keeps each backbone, tokenizer, image processor, and language-model head frozen. Sections 3.1–3.5 specify the repeated-read protocol, attention-ratio computation, token-selection rule, and key-bias injection. We will release the inference implementation, model-specific integration code, evaluation scripts, configuration files, and diagnostic data in an anonymous supplementary package. These materials will provide the information required to reproduce the reported results, subject to the licenses of the underlying models and benchmarks.

Ethics statement

This work introduces an inference-time intervention and uses existing model checkpoints and public evaluation benchmarks. It does not collect personal data or involve human participants. Its downstream risks inherit those of the underlying MLLMs: selectively reinforcing misleading visual evidence may make a confident but incorrect answer more likely. Users should therefore retain task-appropriate verification for consequential deployments.

AI use statement

Generative AI tools were used to assist with manuscript organization, English-language drafting and editing, LaTeX preparation, and code review. The human authors are responsible for the research questions, method, experiments, analysis, interpretation, and final submission. All AI-assisted text, claims, citations, code suggestions, and artifacts were or will be checked by the authors against the source code, experimental records, and primary literature before submission. The authors take responsibility for the final content of this work.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 23716–23736, 2022.

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

Qingqing Cao, Bhargavi Paranjape, and Hannaneh Hajishirzi. Pumer: Pruning and merging tokens for efficient vision language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 12890–12903, 2023.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Proceedings of the European Conference on Computer Vision*, pp. 19–35, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 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, 2024b.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1254–1262, 2024c.

Xiaoxia Cheng, Zeqi Tan, Wei Xue, and Weiming Lu. Information re-organization improves reasoning in large language models. In *Advances in Neural Information Processing Systems*, pp. 130214–130236, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, pp. 49250–49267, 2023.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, pp. 139348–139379, 2024.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 3543–3556, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730–19742, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pp. 34892–34916, 2023.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *ECCV*, pp. 216–233, 2024b.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. In *Advances in Neural Information Processing Systems*, pp. 43447–43478, 2023.

Yi Lu, Jiawang Cao, Yongliang Wu, Bozheng Li, Licheng Tang, Yangguang Ji, Chong Wu, Jay Wu, and Wenbo Zhu. Rsvp: Reasoning segmentation via visual prompting and multi-modal chain-of-thought. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 14699–14716, 2025.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 20023–20070, 2025.

Didac Suris, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11854–11864, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa

Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-Yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. 2025.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *Proceedings of the International Conference on Learning Representations*, pp. 148769–148794, 2026.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 2609–2634, 2023a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations*, 2023b.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems*, pp. 113569–113697, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pp. 24824–24837, 2022.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 11–20, 2019.

Penghao Wu and Saining Xie. 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.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, and Shuai Ma. Re-reading improves reasoning in large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 15549–15575, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. In *Findings of EMNLP*, pp. 11289–11303, 2023.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Advances in Neural Information Processing Systems*, pp. 5168–5191, 2023.

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

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4189–4199, 2025.

Table 3: Sensitivity to the intervention strength α on Qwen2.5-VL-7B. Each fixed α is evaluated on all four benchmarks. Values are percentages; Mean is the macro-average across the four reported metrics. The baseline uses the original one-pass protocol, and $\alpha = 0$ denotes unreinforced Repeat.

α	TreeBench Acc.	WeMath Strict	CharXiv Reasoning	MMBench Acc. (Circ.)	Mean
Original	27.41	36.86	41.20	75.69	45.29
0	29.38	37.90	42.50	79.30	47.27
0.25	28.40	39.05	42.70	79.21	47.34
0.5	30.37	38.67	42.50	79.21	47.69
0.75	29.63	37.43	42.60	79.30	47.24
1	30.86	39.71	42.80	79.30	**48.17**
1.25	31.11	37.71	42.70	79.38	47.73
1.5	32.10	37.14	42.80	79.21	47.81
1.75	32.84	36.19	**43.40**	79.21	47.91
2	**33.09**	35.90	43.00	79.21	47.80

A Implementation Details

All experiments are inference-only and run on NVIDIA A100 GPUs with 80 GB of memory per GPU. We keep the original parameters, tokenizer, image processor, visual encoder, language-model head, rotary-position implementation and query/key normalization of each backbone frozen throughout evaluation. The attention statistics are extracted from the final language-model attention layer; the visual stack and earlier language-model layers retain their native attention implementation. When required by GPU memory, the attention statistics are computed with a chunked online-softmax procedure that retains only the text-to-vision quantities needed by the two reads.

The four benchmarks differ in their data characteristics and evaluation settings. TreeBench contains 405 challenging visual question-answering pairs designed around focused visual perception, traceable visual evidence, and second-order reasoning over object interactions and spatial relations. We-Math contains 6.5K visual mathematics problems spanning 67 hierarchical knowledge concepts and five levels of knowledge granularity; to rigorously evaluate true mathematical comprehension without random guessing, we report both **Strict Accuracy (Strict)**—which requires all sub-questions or variants of a problem to be correctly answered—and **Complete Mastery Strict (CM Strict / CM)**, which demands that a model correctly solves both the foundational single-step sub-problems and the composite multi-step problem. For CharXiv, we evaluate its more challenging Reasoning subset and report Reasoning accuracy. MMBench evaluates fine-grained vision–language abilities through multiple-choice questions and uses Circular Evaluation, where an item with N choices is evaluated under N circularly shifted option arrangements and is counted as correct only when all passes are answered correctly. Together, these benchmarks provide diverse visual evidence structures and evaluation settings for assessing multimodal reasoning.

For every benchmark, we retain the official prompt template and official answer-extraction convention. The same prompt, image preprocessing and decoding rule are used for the one-pass baseline, all repetition ablations and *AMR*. We use greedy decoding with temperature $= 0$. All reported values are percentages under the official benchmark metric definitions.

B Hyperparameter Analysis

We analyze the intervention strength α on Qwen2.5-VL-7B. To isolate the effect of this hyperparameter, each fixed value is applied uniformly to all four benchmarks. The "Original" row is the standard one-pass baseline, and $\alpha = 0$ is the complete ABAB Repeat protocol without additional visual reinforcement. Thus, the sweep separates the gain from repeating the multimodal input from the additional gain due to selective key-bias reinforcement. Table 3 reports the results. The fixed-α macro average increases from 45.29% for Original to 47.27% for unreinforced Repeat, confirming that the repeated multimodal context already provides a substantial gain. Adding visual reinforcement produces its best fixed setting at $\alpha = 1$, with a macro-average of 48.17%. This corresponds to a 2.88-point improvement over Original and a further 0.90-point improvement over $\alpha = 0$.

The per-benchmark trends explain why a moderate shared value is preferable to an aggressively tuned intervention. TreeBench improves progressively for larger values and reaches 33.09% at $\alpha = 2$, but this behavior is not shared by WeMath: WeMath peaks at 39.71% for $\alpha = 1$ and then declines, falling below the Original score at $\alpha = 2$. CharXiv-Reasoning remains positive throughout the sweep and reaches its highest score at $\alpha = 1.75$, whereas MMBench is nearly saturated across the tested range and changes by less than 0.2 points around its maximum. These results indicate that the intervention strength interacts with the evidence demands of each task.

The macro-average nevertheless remains relatively stable for $\alpha \in [0.5, 2]$, ranging from 47.24% to 48.17%. In particular, the performance does not increase monotonically with the bias strength: stronger reinforcement can help TreeBench while suppressing useful evidence on WeMath. The fixed sweep therefore supports using a moderate value selected by aggregate validation rather than choosing the largest value for every task.

The main-table Qwen2.5-VL-7B results may use task-specific settings when reported as the best configuration for an individual benchmark. The conclusion from this appendix concerns the fixed setting used for cross-model evaluation: we adopt $\alpha = 1$ for all models other than those explicitly reported with task-specific Qwen2.5-VL-7B settings. This choice gives the strongest fixed macro-average and avoids making the method depend on a separate per-task search for every backbone.