
LIMITATION, DISCUSSION, AND FUTURE WORK

Beyond a single boundary. Our experiments in Sec. 4.3 and Sec. 4.4 show that the division between perception and reasoning is not determined by a crisp threshold. Instead, performance peaks appear in *bands* of layers, and the optimal ℓ_{perc} and ℓ_{reas} vary across tasks. This suggests that different layers may play mixed roles, and that the model undergoes a gradual transition from perception to reasoning rather than a clean separation. Future work could build on interpretability studies to map these transitions more precisely, moving from a one-dimensional boundary to richer structural patterns.

Toward more adaptive selection. Our method follows the *minimal editing* principle, amplifying a small set of heads without attenuating others. Formally, the gain in task alignment is $\Delta = \sum_{h \in \mathcal{H}} (\gamma_h - 1) u_h$, where u_h measures the usefulness of head h . In this work, we identify perception and reasoning heads using simple ratio thresholds. A natural extension is to design more fine-grained classifiers that score each head by multiple signals (e.g., depth, modality ratio, consistency). An adaptive scheme could then select $S(x)$ depending on the input x , and set $\gamma_h > 1$ only for those heads. While this may yield larger Δ , it also risks higher variance and efficiency costs. Thus, the challenge is to balance stronger, adaptive selection with the stability and efficiency of the current plug-and-play design.

A NOTATIONS AND DEFINITIONS

Table 1: Notations and Definitions

Notation	Definition
$\mathcal{T}_v, \mathcal{T}_t, \mathcal{T}_q$	Sets of vision, text tokens and query tokens, respectively.
l, h	Indices for layer and attention head in a Transformer, respectively.
$X^{(l)}$	Input hidden states at layer l .
$W_Q^{(h,l)}, W_K^{(h,l)}, W_V^{(h,l)}$	Query, key, and value projection matrices for head h in layer l .
$A^{(h,l)}$	Attention matrix for head h in layer l , $A^{(h,l)} \in \mathbb{R}^{N \times N}$.
$a_{ij}^{(h,l)}$	Attention weight from query token i to key token j in matrix $A^{(h,l)}$.
$O^{(h,l)}$	Output vector for head h in layer l , $O_h^{(l)} \in \mathbb{R}^{d_{\text{model}}}$.
$V^{(h,l)}$	Value tensor for head h in layer l .
$Y^{(l)}$	Standard multi-head output of the Transformer block at layer l .
$\mathcal{L}_{\text{perc}}, \mathcal{L}_{\text{reas}}$	Sets of perception layers and reasoning layers, respectively.
$l_{\text{perc}}, l_{\text{reas}}$	Boundaries for the last layer of the perception stage and the first layer of the reasoning stage.
$S_v^{(l)}(h), S_t^{(l)}(h)$	Average attention ratio allocated to visual/textual tokens by head h in layer l .
$\tau_{\text{perc}}, \tau_{\text{reas}}$	Thresholds of visual attention ratio to distinguish perception and reasoning heads.
\mathcal{H}	The full set of attention heads in a given multi-head attention (MHA) sub-layer.
$\mathcal{H}_{\text{perc}}^{(l)}, \mathcal{H}_{\text{reas}}^{(l)}$	Sets of perception-oriented and reasoning-oriented heads identified in layer l .
$g_{\text{perc}}, g_{\text{reas}}$	Global gain factors applied to perception and reasoning heads, respectively.
$\alpha_{\text{perc}}, \alpha_{\text{reas}}$	Scalar gain assigned to a head based on its attention ratio (used interchangeably with $g_{\text{perc}}, g_{\text{reas}}$).
$g^{(h,l)}$	Specific gain value applied to head h in layer l .

Continued on next page

Table 1 – Continued from previous page

Notation	Definition
O_S	Aggregated output of a subset of heads $S \subseteq \mathcal{H}$.

B SELECTIVE ENHANCEMENT AS OUR POLICY CHOICE

Problem and Objective (Minimal Editing Principle). When intervening on attention heads, our objective is to *maximize the correction of hallucination* while *minimizing unnecessary edits* to the model’s internal representations. We formalize this idea as the **Minimal Editing Principle**: an effective intervention should (1) amplify signals that are already verified as beneficial, and (2) avoid suppressing or perturbing heads whose functions remain uncertain.

Let \mathcal{H} denote the full set of attention heads in a given multi-head attention (MHA) sub-layer. Each head $h \in \mathcal{H}$ produces an output vector $\mathbf{O}_h^{(l)} \in \mathbb{R}^{d_{\text{model}}}$, and the aggregated MHA output can be expressed as

$$\tilde{Y}^{(l)} = \sum_{h \in \mathcal{H}} \gamma_h^{(l)} \mathbf{O}_h^{(l)}, \quad (1)$$

where $\gamma_h^{(l)}$ is a multiplicative gain applied to head h . In this formulation, any intervention corresponds to editing the gain pattern $\{\gamma_h^{(l)}\}_{h \in \mathcal{H}}$. *Minimal editing* requires that such modifications keep the original contributions of most heads, while selectively reinforcing those aligned with perception or reasoning.

Intuitively, this principle reflects a conservative stance. We cannot guarantee that non-target heads are harmful. If they are attenuated indiscriminately, the model may lose useful functions and suffer collateral damage. The safer strategy is therefore to intervene as little as possible: amplify only the identified functional heads, while keeping the rest unchanged to preserve the existing representation.

Formal Setup and Minimal Assumptions. To make analysis tractable and verifiable, we adopt two minimal assumptions: (1) interventions operate *after attention computation but before residual and normalization*, ensuring that changes are linear and directly comparable; (2) all enhanced heads share a common amplification factor $\alpha > 1$, while all attenuated heads (if any) share a common factor $\beta < 1$.

These assumptions preserve the generality of our formulation while guaranteeing that differences between strategies can be derived in closed form. In the next subsection, we instantiate four policies under this setup and analyze their functional differences.

Four Strategies and a Core Proposition. Under the above setup, we distinguish four representative policies for modifying head gains:

- **Strategy A (Selective Enhancement).**

$$\gamma_h^{(l)} = \begin{cases} \alpha, & h \in \mathcal{H}_{\text{enhance}}, \\ 1, & h \in \mathcal{H} \setminus \mathcal{H}_{\text{enhance}}. \end{cases} \quad (2)$$

- **Strategy B (Selective Attenuation).**

$$\gamma_h^{(l)} = \begin{cases} \beta, & h \in \mathcal{H}_{\text{attenuate}}, \\ 1, & h \in \mathcal{H} \setminus \mathcal{H}_{\text{attenuate}}. \end{cases} \quad (3)$$

- **Strategy C (Bipolar Scaling).**

$$\gamma_h^{(l)} = \begin{cases} \alpha, & h \in \mathcal{H}_{\text{enhance}}, \\ \beta, & h \in \mathcal{H}_{\text{attenuate}}. \end{cases} \quad (4)$$

- **Strategy D (Mixed Policy).**

$$\gamma_h^{(l)} = \begin{cases} \alpha, & h \in \mathcal{H}_{\text{enhance}}, \\ \beta, & h \in \mathcal{H}_{\text{attenuate}}, \\ 1, & h \in \mathcal{H}_{\text{neutral}}. \end{cases} \quad (5)$$

Proposition (Difference Equation). Let $\mathbf{O}_S = \sum_{h \in S} \mathbf{O}_h^{(l)}$ denote the aggregated output of a subset of heads $S \subseteq \mathcal{H}$. Then, relative to Strategy A, any scheme that introduces attenuation (C or D) differs by an explicit subtractive term:

$$\tilde{Y}_C^{(l)} - \tilde{Y}_A^{(l)} = (\beta - 1) \mathbf{O}_{\mathcal{H}_{\text{att}}}, \quad (6)$$

$$\tilde{Y}_D^{(l)} - \tilde{Y}_A^{(l)} = (\beta - 1) \mathbf{O}_{\mathcal{H}_{\text{att}}}, \quad (7)$$

where \mathcal{H}_{att} denotes the set of attenuated heads.

Proof. Both Strategy A and Strategy C/D share the amplified component $\alpha \mathbf{O}_{\mathcal{H}_{\text{enh}}}$. The difference arises in the treatment of non-enhanced heads: Strategy A preserves their outputs unchanged, while C and D rescale them by $\beta < 1$. This introduces a correction term $(\beta - 1) \mathbf{O}_{\mathcal{H}_{\text{att}}}$, representing the explicit removal of information contributed by attenuated heads.

Interpretation. This difference equation makes the design trade-off explicit: ♣ Strategy A amplifies useful heads while keeping all others intact. ♣ Strategies C and D apply an additional subtraction to all attenuated heads, implicitly assuming they are harmful. Since non-target heads may still carry latent but beneficial functions, attenuation risks collateral degradation. Thus, the *minimal editing* choice is Strategy A, which strengthens the identified heads without disturbing the rest of the model.

Expectation-Level Intuition: Why Selective Enhancement is Safer. Consider a task-aligned direction $v \in \mathbb{R}^{d_{\text{model}}}$, and let each head’s contribution be

$$u_h := \langle \mathbf{O}_h^{(l)}, v \rangle, \quad (8)$$

which measures how much the head h aligns with the task. For any headset S , we then define its total contribution as

$$U(S) = \sum_{h \in S} u_h. \quad (9)$$

Ignoring LayerNorm rescaling, the alignment change caused by a gain pattern $\{\gamma_h^{(l)}\}$ is

$$\Delta(\{\gamma\}) \approx \sum_{h \in \mathcal{H}} (\gamma_h^{(l)} - 1) u_h, \quad (10)$$

which simply adds the weighted shifts of each head relative to the neutral baseline ($\gamma_h = 1$).

Three Strategies. Applying this formulation, the alignment increments become:

$$\Delta_A \approx (\alpha - 1) U(\mathcal{H}_{\text{enhance}}), \quad (11)$$

corresponding to *selective enhancement*, where only useful heads are amplified.

$$\Delta_B \approx (\beta - 1) U(\mathcal{H}_{\text{attenuate}}), \quad (12)$$

for *selective attenuation*, where some heads are suppressed while all others remain neutral.

$$\Delta_C \approx (\alpha - 1) U(\mathcal{H}_{\text{enhance}}) + (\beta - 1) U(\mathcal{H}_{\text{attenuate}}), \quad (13)$$

for *bipolar scaling*, where enhancement and attenuation are applied simultaneously.

Assumption (non-harmful heads). Prior interpretability studies indicate that most attention heads carry diverse or weakly positive roles rather than being systematically harmful (Olsson et al., 2022). Formally, this suggests

$$\mathbb{E}[U(\mathcal{H}_{\text{attenuate}})] \geq 0, \quad (14)$$

meaning that, on average, attenuated heads are not expected to reduce alignment.

Implications. Taking expectations under this assumption yields

$$\mathbb{E}[\Delta_A] > 0, \quad \mathbb{E}[\Delta_B] \leq 0, \quad \mathbb{E}[\Delta_C] \leq \mathbb{E}[\Delta_A]. \quad (15)$$

Hence, Strategy A provides a reliably positive gain by amplifying identified functional heads. Strategies B and C both introduce subtractive terms $(\beta - 1) U(\mathcal{H}_{\text{attenuate}})$, which are non-positive in expectation, making them less stable. This formalizes why, under the *minimal-editing* principle, selective enhancement is the safer choice.

Takeaways. Strategy A yields a guaranteed positive gain by enhancing only functional heads. Strategies B and C involve attenuation terms $(\beta - 1) U(\mathcal{H}_{\text{attenuate}})$, which are non-positive in expectation and risk weakening neutral or beneficial heads. Therefore, under the *minimal editing* principle, Strategy A is the most stable and reliable choice.

C TIME COMPLEXITY ANALYSIS

In this section, we present a detailed analysis of the algorithmic time complexity of the standard attention mechanism and our proposed method in a single multi-head attention block.

Let the input sequence length be N , the model hidden dimension be d_{model} , and the number of attention heads be H . For each attention head, the query, key, and value dimensions are denoted by d_k and d_v , with $d_k = d_v = d_{model}/H$. For notational simplicity, we refer to the per-head dimension as d_h . In the following analysis, we focus on the asymptotic complexity with respect to N , treating d_{model} , H , d_k , and d_v as constants.

STANDARD ATTENTION

Input Projections. Given the input tensor $X \in \mathbb{R}^{N \times d_{model}}$, three linear projections are applied to generate the query, key, and value tensors: $Q = XW_Q$, $K = XW_K$, and $V = XW_V$, where $W_Q, W_K, W_V \in \mathbb{R}^{d_{model} \times d_{model}}$. Each projection is a matrix multiplication with complexity $O(N \cdot d_{model}^2)$, resulting in:

$$T_1 = O(3 \cdot N \cdot d_{model}^2) = O(N \cdot d_{model}^2) \quad (16)$$

Attention Score Computation. For each head $h \in \{1, \dots, H\}$, unnormalized attention scores are computed as $Q_h K_h^T$, where $Q_h \in \mathbb{R}^{N \times d_k}$ and $K_h^T \in \mathbb{R}^{d_k \times N}$. The cost per head is $O(N^2 \cdot d_k)$, and across all heads:

$$T_2 = O(H \cdot N^2 \cdot d_k) \quad (17)$$

Softmax Normalization. A row-wise softmax is applied to the $N \times N$ score matrix for each head. Applying softmax to a vector of length N costs $O(N)$, and thus to N rows is $O(N^2)$ per head. Across all heads:

$$T_3 = O(H \cdot N^2) \quad (18)$$

Value Aggregation. The normalized attention matrix A_h is multiplied by $V_h \in \mathbb{R}^{N \times d_v}$ to produce the head output $O_h \in \mathbb{R}^{N \times d_v}$. The cost per head is $O(N^2 \cdot d_v)$, yielding:

$$T_4 = O(H \cdot N^2 \cdot d_v) \quad (19)$$

Output Projection. The H head outputs are concatenated into a $N \times d_{model}$ matrix and projected via $W_O \in \mathbb{R}^{d_{model} \times d_{model}}$ with cost:

$$T_5 = O(N \cdot d_{model}^2) \quad (20)$$

Summing all terms gives the total time complexity:

$$T_{\text{Standard}} = T_1 + T_2 + T_3 + T_4 + T_5 = O(Nd_{model}^2 + HN^2d_k + HN^2 + HN^2d_v + Nd_{model}^2)$$

As N grows, the N^2 terms dominate. Therefore, the overall asymptotic time complexity is:

$$T_{\text{Standard}} = O(H \cdot N^2 \cdot (d_k + d_v + 1)) = O(N^2) \quad (21)$$

OUR METHOD

Standard Attention Score and Softmax Computation. Our method first performs the standard attention computation as in the baseline, obtaining all H normalized attention matrices $A_h \in \mathbb{R}^{N \times N}$:

$$T'_1 = O(Nd_{model}^2 + HN^2d_k + HN^2) \quad (22)$$

Visual Attention Ratio Calculation. We then compute a visual attention ratio for each head. This operation traverses all $N \times N$ attention entries, giving:

$$T'_2 = O(H \cdot N^2) \quad (23)$$

Head Categorization and Gain Assignment. Each head is categorized and assigned a scalar gain α_h based on its $S_v(h)$. This involves only $O(H)$ comparisons and assignments, which are negligible compared to the dominant terms.

Table 2: Average per-batch inference time (seconds). Parentheses show relative change vs. Vanilla.

Method	MathVista _{mini}	MathVision _{mini}	HallusionBench	MMStar	SEED-Bench
Vanilla	98 (+0.0%)	98 (+0.0%)	101 (+0.0%)	83 (+0.0%)	68 (+0.0%)
VCD	159 (+62.2%)	241 (+145.9%)	170 (+68.3%)	113 (+36.1%)	98 (+44.1%)
CGD	482 (+391.8%)	709 (+623.5%)	664 (+557.4%)	297 (+257.8%)	329 (+383.8%)
AGLA	119 (+21.4%)	164 (+67.3%)	123 (+21.7%)	87 (+4.8%)	88 (+29.4%)
Ours	103 (+5.1%)	101 (+3.1%)	103 (+2.0%)	83 (+0.0%)	69 (+1.5%)

Modulated Value Aggregation. The values are aggregated as $O_h = A_h V_h$, followed by element-wise scaling with α_h . The aggregation part has cost:

$$T'_{4a} = O(H \cdot N^2 \cdot d_v) \quad (24)$$

and the scaling part adds $O(H \cdot N \cdot d_v) = O(N \cdot d_{model})$, which is dominated by the aggregation cost. Thus:

$$T'_4 = O(H \cdot N^2 \cdot d_v) \quad (25)$$

Final Output Projection. The modulated outputs are concatenated and projected as in the baseline:

$$T'_5 = O(N \cdot d_{model}^2) \quad (26)$$

Combining all steps, the total complexity of our method is:

$$T_{\text{Ours}} = T'_1 + T'_2 + T'_4 + T'_5 = O(N d_{model}^2 + H N^2 d_k + 2 H N^2 + H N^2 d_v + N d_{model}^2)$$

Focusing on the dominant terms with respect to N , we obtain:

$$T_{\text{Ours}} = O(H \cdot N^2 \cdot (d_k + d_v + 2)) = O(N^2) \quad (27)$$

EXPERIMENTS

As reported in Table 2, our method attains per-batch inference times that are essentially indistinguishable from the Vanilla baseline across all six benchmarks, with only marginal fluctuations of a few percentage points. This empirical parity aligns with the theoretical expectation that our additional head-level operations introduce only constant-factor overhead beyond standard multi-head attention. Consequently, the overall time complexity remains unchanged, and in practice, the runtime of our approach matches that of the original model.

C.1 COMPARISON AND CONCLUSION

Although our method introduces an additional $O(H \cdot N^2)$ step for head-level ratio computation, the total complexity remains dominated by the $O(N^2)$ cost of the standard attention mechanism. As a result, both exhibit the same asymptotic time complexity of $O(N^2)$ with respect to the sequence length N . The additional operations only contribute constant-factor overhead without altering the overall computational order.

Consequently, in practice, the runtime of our approach is **essentially indistinguishable** from that of standard attention.

D IMPLEMENTATION DETAILS

Ocean-R1 (Ming et al., 2025) Ocean-R1 is a large vision-language reasoning model fine-tuned from Qwen2.5-VL-Instruct (Bai et al., 2025), designed to enhance cross-modal reasoning and visual understanding capabilities through a two-stage rule-based Reinforcement Learning framework. The first stage focuses on strengthening the model’s reasoning ability, while the second stage improves its visual perception. Experimental results demonstrate that Ocean-R1 achieves substantial

Table 3: Selected Hyperparameters

Model	ℓ_{reas}	τ_{reas}	g_{reas}	ℓ_{perc}	τ_{perc}	g_{perc}
Kimi-VL-A3B-Thinking	5	0.01	1.40	10	0.27	1.20
Ocean-R1-7B-Instruct	3	0.01	1.30	7	0.22	1.16
R1-Onevision	3	0.01	1.30	7	0.30	1.20

performance gains, particularly on visual mathematical reasoning benchmarks such as MathVision (+2.7/+2.7) and MathVista (+4.9/+4.4), showing strong multimodal reasoning and generalization capabilities.

R1-Onevision (Yang et al., 2025) R1-Onevision is a state-of-the-art multimodal reasoning model designed to bridge the gap between visual perception and deep reasoning. It is fine-tuned from Qwen2.5-VL (Bai et al., 2025), with a focus on cross-modal reasoning that enables precise understanding and processing of both visual and textual information. Unlike previous models that primarily rely on fixed structures for reasoning, R1-Onevision employs a two-stage post-training strategy: Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL), enhancing its ability to generalize across diverse tasks. The model leverages a cross-modal reasoning pipeline that transforms images into formal text-based representations, which are then processed to generate structured reasoning paths. It also incorporates a "role-playing" strategy to iteratively refine visual comprehension, ensuring robust multimodal coherence. Experimental evaluations on benchmarks like MathVista and MathVerse demonstrate that R1-Onevision outperforms several state-of-the-art models, including GPT-4o and Qwen2.5-VL, showcasing its superior reasoning and generalization capabilities.

Kimi-VL (Team et al., 2025) Kimi-VL is an efficient open-source vision-language model built upon a Mixture-of-Experts (MoE) language decoder with only 2.8B activated (16B total) parameters, paired with a 400M native-resolution vision encoder (MoonViT). It is designed to provide advanced multimodal reasoning, long-context understanding, and strong agent capabilities, while maintaining high parameter efficiency. Unlike most dense-architecture VLMs, Kimi-VL achieves competitive or superior performance to larger models on diverse tasks, including college-level problem solving, OCR, multi-image reasoning, video understanding, and long-document comprehension. Through long Chain-of-Thought supervised fine-tuning and reinforcement learning, its enhanced variant Kimi-VL-Thinking demonstrates strong long-horizon multimodal reasoning ability, achieving remarkable results on benchmarks such as MathVision and MathVista. This demonstrates Kimi-VL’s effectiveness in combining parameter efficiency with powerful multimodal reasoning capabilities.

Our method is primarily implemented in the `eager_attention_forward` function within the modeling file (`modeling_qwen2_5_vl.py` and `modeling_kimi_vl.py`). Also, we incorporate a caching mechanism to store essential information (e.g., the range of visual tokens and hyperparameters).

Table 3 summarizes the hyperparameters used in our experiments.

E DATASETS

MathVista (Lu et al., 2023) Aggregates 6,141 problems by consolidating 28 existing multimodal datasets with three newly curated sources—IQTest (puzzle figures), FunctionQA (function plots), and PaperQA (figures from academic papers)—spanning charts, diagrams, textbook geometry, and everyday VQA scenes. It targets cross-domain mathematical reasoning grounded in images, with compositional perception and minimal leakage, and remains challenging for current models.

MathVision (Wang et al., 2024) Curated from real math competitions to provide 3,040 image-based problems across 16 mathematical disciplines (e.g., analytic geometry, topology, graph theory) and five difficulty levels. Designed to stress rigorous multimodal mathematical reasoning beyond earlier benchmarks, it exhibits a large human–model gap and offers *test* and *test-mini* splits for rapid benchmarking.

MMStar (Chen et al., 2024) Built by screening items from existing benchmarks and then manually vetting them to ensure vision indispensability and remove data leakage, yielding 1,500 human-selected samples covering six core capabilities along 18 axes. Its goal is a purified, balanced test that measures true multimodal gains—i.e., questions where visual evidence is necessary and difficulty stems from advanced cross-modal reasoning.

HallusionBench (Guan et al., 2024) Comprises 346 figures from diverse sources and formats (e.g., charts, tables, maps, famous optical illusions, memes) paired with 1,129 expert-authored questions structured into control pairs and “visual-dependent” vs. “visual-supplement” settings. It is purpose-built to disentangle language hallucination from visual illusion and to test consistency under easy/hard conditions, remaining challenging for state-of-the-art LVLMs.

SEED-Bench (Li et al., 2024) Uses human-verified multiple-choice questions targeted to capability dimensions; in SEED-Bench-1 (19k), images come from Conceptual Captions and videos from Something-Something V2, Epic-Kitchens, and Breakfast, covering both spatial and temporal understanding. The benchmark’s design goal is dimension-specific difficulty and objective automatic scoring across 12 image/video comprehension dimensions.

F BASELINES

We adopt a deliberately diverse set of baselines. Methods for mitigating inference-time hallucinations via generation intervention can be grouped into three families: **Contrastive Decoding**, **Guided Decoding**, and **Visual Amplification**. Accordingly, we select VCD, CGD, and AGLA as representative instances. These baselines instantiate three complementary mechanisms—probability calibration, external guidance, and feature enhancement—covering different failure modes while reducing evaluation bias due to methodological homogeneity; together they constitute a representative comparison suite.

Visual Contrastive Decoding (VCD) (Leng et al., 2024). VCD is a training-free decoding strategy that explicitly builds a counterfactual view of the image by introducing controlled visual uncertainty and then contrasts the model’s predictions under the original versus the uncertain view. Tokens that are preferred only when the image is ambiguous are down-weighted, which directly combats two root causes of object hallucination: over-reliance on language priors and spurious object co-occurrence. An adaptive plausibility constraint further preserves fluent generation by truncating to high-confidence candidates from the original distribution. The key difference from other baselines is that VCD does not require external models or retraining; it modifies only the sampling distribution and therefore generalizes across LVLm families.

Assembly of Global and Local Attention (AGLA) (An et al., 2025). AGLA targets the attention deficiency behind many hallucinations: LVLms tend to lock onto prompt-irrelevant global patterns while missing fine-grained, prompt-relevant regions. It first performs image–prompt matching (via a Grad-CAM–style analysis) to produce an augmented view that highlights regions tied to the query and suppresses distractors; decoding then fuses evidence from the global view (original image) and the local, discriminative view (augmented image). This design differs from VCD and CGD by enriching the *visual features* themselves rather than only reshaping probabilities, which improves grounding without sacrificing generative context. Empirically, it is especially strong under challenging negative settings, e.g., achieving about 81.36 F1 on the adversarial split of POPE with LLaVA-1.5.

CLIP-Guided Decoding (CGD) (Deng et al., 2024). CGD injects an external, vision-language prior—CLIP—directly into decoding, but at the *sentence level*. Instead of judging partial tokens, CGD scores complete sentence candidates by mixing the LVLm’s length-normalized likelihood with CLIP image–text similarity and keeps candidates that are both probable and visually grounded. Two design choices explain its gains: (i) sentence-level guidance avoids the myopia of token-level heuristics and directly penalizes late-caption drift (later sentences are empirically more hallucinatory), and (ii) CLIP similarity provides a stronger, more stable signal of image–text alignment than model likelihood alone. As a result, CGD substantially lowers hallucination while preserving caption quality, e.g., CHAIRs on COCO with LLaVA-1.5 drops from 44.7 to 29.7.

G METRICS

We adopt the *F1 score* as the primary evaluation metric, with calculation logic adapted to the nature of each dataset. Specifically, two distinct evaluation schemes are applied: (1) a binary judgment scheme for datasets containing binary ground-truth labels (e.g., HallucinationBench), and (2) a multi-class scheme for datasets containing multiple-choice questions (e.g., MMStar and SEED-Bench). In both schemes, the final performance is reported using the *Weighted F1 score*.

Binary Judgment Scheme For binary datasets, we use the standard F1 score, defined as:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}. \quad (28)$$

Multi-class Scheme For multi-class datasets, each sample has a ground-truth label $y \in \{1, \dots, K\}$ and a predicted label $\hat{y} \in \{1, \dots, K\}$. For each class $i \in \{1, \dots, K\}$, we define:

$$TP_i = |\{\hat{y} = i, y = i\}|, \quad FP_i = |\{\hat{y} = i, y \neq i\}|, \quad (29)$$

$$FN_i = |\{\hat{y} \neq i, y = i\}|, \quad TN_i = |\{\hat{y} \neq i, y \neq i\}|. \quad (30)$$

Then:

$$\text{Precision}(i) = \frac{TP_i}{TP_i + FP_i}, \quad \text{Recall}(i) = \frac{TP_i}{TP_i + FN_i}, \quad (31)$$

$$F1(i) = \frac{2 \times \text{Precision}(i) \times \text{Recall}(i)}{\text{Precision}(i) + \text{Recall}(i)}. \quad (32)$$

Weighted Aggregation To account for class imbalance, the weighted F1 score is computed as:

$$\text{Weighted-F1} = \sum_{i=1}^K \frac{\text{Support}(i)}{N} \times F1(i), \quad (33)$$

where $\text{Support}(i)$ denotes the number of samples with ground-truth label i , and N is the total number of samples.

H CASE ANALYSIS

H.1 CONTRIBUTION MAP

Following Michel et al (Michel et al., 2019), we quantify the contribution of each attention head to a *specific next-token prediction* by inserting differentiable gates on head outputs and reading out the loss gradients w.r.t. these gates. Intuitively, if slightly amplifying head (l, h) would decrease the loss for the current target token, that head is *helpful* for this token; if it would increase the loss, it is *harmful*. We report absolute sensitivities as non-negative importance scores and optionally keep the signed scores to analyze supportive vs. adversarial heads.

Notation at a glance.

- B : batch size; S : sequence length seen by the model at the current decoding step (includes prompt and previously generated tokens).
- L : number of transformer layers.
- H : number of attention heads per layer.
- d_{model} : hidden size; $d_{\text{head}} = d_{\text{model}}/H$: per-head width after splitting.
- \mathbf{x} : model inputs (encoded sequence including both image and text); $\mathbf{y}_{<\tau}$: generated prefix; y_τ : random variable at current position; t^* : target token for this step.
- $\mathcal{L} = -\log P(y_\tau = t^* | \mathbf{x}, \mathbf{y}_{<\tau})$: cross-entropy loss defined only for current position τ .

H.1.1 GATE PARAMETER MECHANISM

We insert lightweight gate parameters onto each attention head to directly modulate their outputs without changing model weights. This setup allows us to treat each gate as a differentiable scalar that controls the influence of its head. With this mechanism in place, we can measure how scaling a head affects the model’s loss for the current token.

For layer l , let the standard multi-head output (after attention and output projection, before the MLP) be

$$\mathbf{O}^{(l)} \in \mathbb{R}^{B \times S \times d_{\text{model}}}. \quad (34)$$

We reshape to expose heads:

$$\mathbf{O}_{\text{heads}}^{(l)} = \text{reshape} \left(\mathbf{O}^{(l)}, [B, S, H, d_{\text{head}}] \right). \quad (35)$$

We use a trainable-on-the-fly (but *not optimized*) gate vector $\mathbf{g}^{(l)} \in \mathbb{R}^H$ initialized as $\mathbf{1}_H$ and broadcast it multiplicatively:

$$\forall b \in [B], s \in [S], h \in [H] : \quad \mathbf{O}_{\text{gated}}^{(l)}[b, s, h, :] = g_h^{(l)} \mathbf{O}_{\text{heads}}^{(l)}[b, s, h, :]. \quad (36)$$

Finally we fold heads back to d_{model} and continue the standard forward. Implementation uses lightweight forward hooks; model weights remain frozen. Gates are created with `requires_grad=True` so that backprop can produce $\frac{\partial \mathcal{L}}{\partial g_h^{(l)}}$.

Shape sanity check. $\mathbf{O}_{\text{heads}}^{(l)}[b, s, h, :] \in \mathbb{R}^{d_{\text{head}}}$ is the contribution of head h at token position s . Multiplying by $g_h^{(l)}$ scales the entire vector produced by that head uniformly at all positions in the current pass; this matches the "masking/scale" abstraction in Michel et al.

H.1.2 GRADIENT-BASED IMPORTANCE

By backpropagating the cross-entropy loss through these gates, we obtain per-head gradients that quantify how loss would change if a head were amplified. Negative gradients indicate supportive heads, while positive gradients reveal harmful ones. Taking the absolute values yields stable importance scores for both visualization and analysis.

At decoding step τ with target t^* ,

$$\mathcal{L} = -\log P(y_\tau = t^* \mid \mathbf{x}, \mathbf{y}_{<\tau}). \quad (37)$$

We define the *signed sensitivity* and the *importance* for head (l, h) :

$$S^{(l,h)} = \frac{\partial \mathcal{L}}{\partial g_h^{(l)}}, \quad I^{(l,h)} = \left| S^{(l,h)} \right|. \quad (38)$$

Interpretation: a negative $S^{(l,h)}$ means "if we scale up this head, loss goes down" (supportive head); a positive $S^{(l,h)}$ suggests the opposite (potentially harmful). Reporting $|S|$ yields non-negative "importance heatmaps", while retaining the sign of S allows for more fine-grained adversarial analysis.

Connection to first-order Taylor pruning. Expanding \mathcal{L} at $\mathbf{g} = \mathbf{1}$, a small perturbation $\Delta g_h^{(l)}$ changes the loss by $\approx S^{(l,h)} \Delta g_h^{(l)}$; therefore $|S^{(l,h)}|$ can directly serve as a "influence measure for current sample/step", which is the per-sample instantiation of the expected sensitivity $I_h = \mathbb{E} |\partial \mathcal{L} / \partial \xi_h|$ proposed by Michel et al. under our gate parameterization.

H.1.3 FROM PER-STEP SCORES TO A RANKING

Finally, we aggregate these per-head importance scores and normalize them either layer-wise or globally. Sorting the normalized scores produces a ranking that highlights which heads are most influential for the given token (or across tokens when averaged). This ranking provides a principled basis for head attribution and comparison.

Collect raw scores

$$\mathcal{I} = \{I^{(l,h)} \mid l = 1, \dots, L; h = 1, \dots, H\}. \quad (39)$$

We consider two normalizations:

Layer-wise (robust to layer-scale differences):

$$\tilde{I}_{\text{layer}}^{(l,h)} = \frac{I^{(l,h)}}{\sum_{h'=1}^H I^{(l,h')}}. \quad (40)$$

Global (for global ranking across the whole stack):

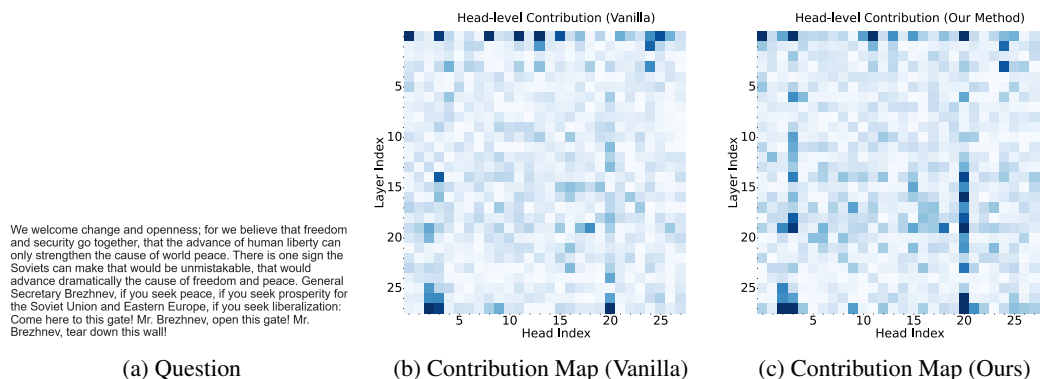
$$\tilde{I}_{\text{global}}^{(l,h)} = \frac{I^{(l,h)}}{\sum_{l'=1}^L \sum_{h'=1}^H I^{(l',h')}}. \quad (41)$$

We then sort

$$\pi = \text{argsort}(\tilde{I}, \text{descending}), \quad (42)$$

obtaining a head ranking tailored to the current token t^* (or to an aggregate of tokens; see below).

H.2 CASE I: REASONING DRIFT



Question - Image (a)

Is the text in this figure trying to talk to Gorbachev?

Vanilla - Image (b)

<think>

Okay, so I need to figure out whether the text in the image is trying to talk to Gorbachev. Let's start out by understanding the context. The image shows a text message or a letter. The text is talking about change and openness and mentions General Secretary Brezhnev. So, Brezhnev was the leader of the Soviet Union during the Cold War. Gorbachev came later, so he might have been the General Secretary later in the 80s.

Looking at the text, it says things like "Come here to this gate!" and "Tear down this wall!". These phrases sound like they're addressing someone specific, probably someone in power authority, which would be the General Secretary. The image also has uses the term "General Secretary Brezhneve", which is a typo for Brezhnev. This reinforces the idea that the speaker is talking to someone in authority.

Also, the message is urgent to to the cause of freedom and peace and economic prosperity in Eastern Europe. This kind of plea would be directed at someone who is in power, like a leader. **Gorbachev** was famous for bringing Soviet policies and for making reforms like glasuez, which aimed at openness and change in the Soviet Union. So, it makes sense that

this message is addressed to him.

The image shows a letter by someone urging change and openness, asking Brezhnev to do something, and even mentioning tearing down a wall. All these points to in message intended at **Gorbachev**. The reference to Brezhnev being the leader, even if the name spelling is incorrect, points to the main target of the message. So, **yes, the text is trying to talk to Gorbachev**.

</think>

The text in the image is clearly directed at Gorbachev. The mention of General Secretary Brezhnev, the reference to tearing down a wall, and the plea for change and openness all point to in address to Gorbachev, who a leader known for the reforms aimed at opening in Eastern Europe. Therefore, the message is addressed to Gorbachev.

****Answer:** Yes, the text in the image is trying to talk to Gorbachev.**

Ours - Image (c)

<think>

Okay, so I'm trying to determine if the text is trying to talk to Gorbachev. Let's start with the image directly.

The image shows a request to open the gates. However, the text in the image is addressed to General Secretary Brezhnev. The image itself depicts a scene where Gorbachev might be the leader of the Soviet Union. The image also a reference to " Gorbachev, which gives a hint that the leader might Gorbachev's the leader of the Soviet Union.

The image shows a request to open the gates. The text in the image is addressed to General Secretary Brezhnev. The image itself shows a scene where Gorbachev might be the leader of the Soviet Union. The image also refers to Gorbachev, but the text is addressed to Brezhnev.

Therefore, the text is not directly addressing Gorbachev; it is instead addressed to Brezhnev, the leader of the Soviet Union at the time.

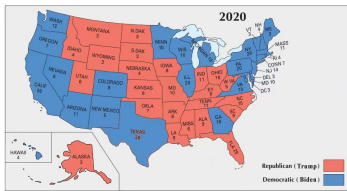
</think>

The image shows a request to open the gates. Therefore, the text is not trying to talk to Gorbachev; it is trying to talk to Brezhnev, the leader of the Soviet Union at the time.

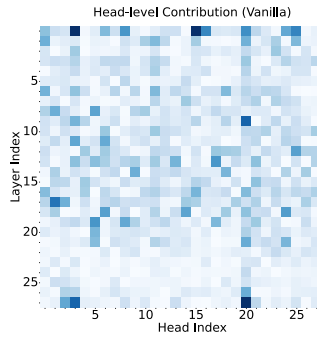
****Answer:** No, the text in the image is not trying to talk to Gorbachev; it is referring to Brezhnev, the leader of the Soviet Union at the time.**

Analysis. From the contribution maps, our method markedly amplifies the mid-to-deep *reasoning heads* in Layer 10–23, thereby substantially improving the model's reasoning capacity. In contrast, the Vanilla model shows no obvious perceptual failure but suffers during reasoning: because the question mentions "Gorbachev", Vanilla model erroneously binds this cue to the entire message, accumulates evidence under the hypothesis "target = Gorbachev," and drifts away from the correct logical chain, ultimately misclassifying an appeal addressed to Brezhnev as speaking to Gorbachev. By strengthening these mid-to-deep reasoning heads, our method maintains stable entity grounding and a consistent causal narrative, **preventing reasoning drift** and producing the correct negative answer.

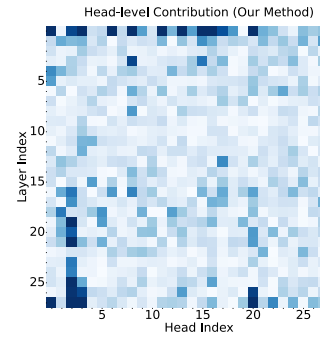
H.3 CASE II: PERCEPTUAL BIAS



(a) Question



(b) Contribution Map (Vanilla)



(c) Contribution Map (Ours)

Question - Image (a)

Based on the map, did the Republic Party win Texas in the 2020 elections?

Vanilla - Image (b)

<think>

Okay, so I'm trying to determine whether the Republican Party won Texas in the 2020 elections based on the provided map. Let me break this down step.

First, the image shows a map of the United States indicating the election results by state for the 2020 U.S. presidential election. Each state is colored either **red** or blue, representing Republican and Democratic wins, respectively. The image shows Texas colored in **red**.

To confirm, the image legend has the **red** color for Republican (Trump) and blue for Democratic (Biden). Therefore, since Texas is colored **red**, it indicates that the Republican Party won Texas in the 2020 election.

There are no other states mentioned that would affect this conclusion. The other states might be colored differently, but Texas being **red** is the key point of this question.

So, putting it all, the image directly shows Texas as **red**, which means the Republican Party won Texas in the 2020 election.

</think>

The image shows Texas colored **red**, which indicates that the Republican Party won Texas in the 2020 election.

Answer: **Yes, the Republican Party won Texas in the 2020 election.**

Ours - Image (c)

<think>

Okay, so I'm trying to determine whether the Republican Party won Texas in the 2020 elections based on the provided map. Let me go through this step by step.

First, I need to understand what the map is showing. It is an electoral map from the 2020 U.S. Presidential Election, where each state is colored either **red** or blue. **Red** represents Republican candidates, and blue represents Democratic candidates. Each state also have a number 2020, so this is the 2020 election results.

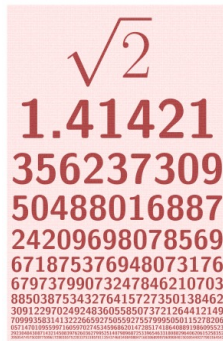
Next, I notice that Texas is colored **blue**. This indicates that Texas was won by the Democrats. Therefore, the Republican Party did not win Texas in the 2020 elections.

</think>

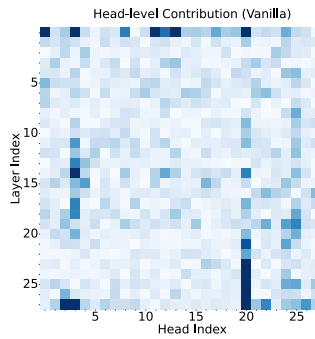
The answer is that the Republican Party did not win Texas in the 2020 elections. Texas was won by the Democrats.

Analysis. In this case, the contribution maps reveal that our method effectively mitigates the dispersed and unfocused attention observed in the Vanilla model by strengthening both shallow perception heads and deep reasoning heads. The primary issue with the Vanilla model arises in the second paragraph of its reasoning, where it misperceives the color of Texas as red. This initial perceptual bias propagates through its logical reasoning process, ultimately leading to an incorrect conclusion that the Republican Party won Texas. In contrast, our method, supported by **stronger perceptual attention** in shallow layers and more robust reasoning capacity in deeper layers, accurately perceives Texas as blue and correctly infers that the state was won by the Democrats. This demonstrates how our approach enhances the interplay between perception and reasoning, enabling the model to avoid compounding errors and to arrive at the correct answer.

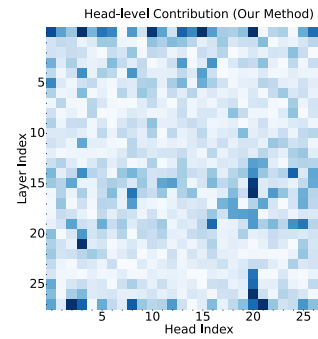
H.4 CASE III: COMPREHENSIVE HALLUCINATION



(a) Question



(b) Contribution Map (Vanilla)



(c) Contribution Map (Ours)

Question - Image (a)

According to the image, does the value of the square root of 2 range from 1.40 to 1.41?

Vanilla - Image (b)

<think>

Okay, so I need to determine whether the square root of 2 ranges from 1.40 to 1.41 based on the image provided. The image shows **several** square roots, and I need to identify the square root of 2.

1. As the image shows **several** square roots, and the first one 2 is approximately **1.44**.
2. The image also shows the square root of 2 as approximately **1.44**.
3. Therefore, the image shows the square root of 2 as approximately **1.44**.

</think>

Answer: The square root of 2 is not **between 1.0 and 1.1**; it is approximately 1.4.

Ours - Image (c)

<think>

Okay, so I need to figure out whether the value of the square root of 2, denoted as $\text{sqrt}(2)$, falls within the range of 1.40 to 1.41. Let's start directly based on the image and the information

provided.

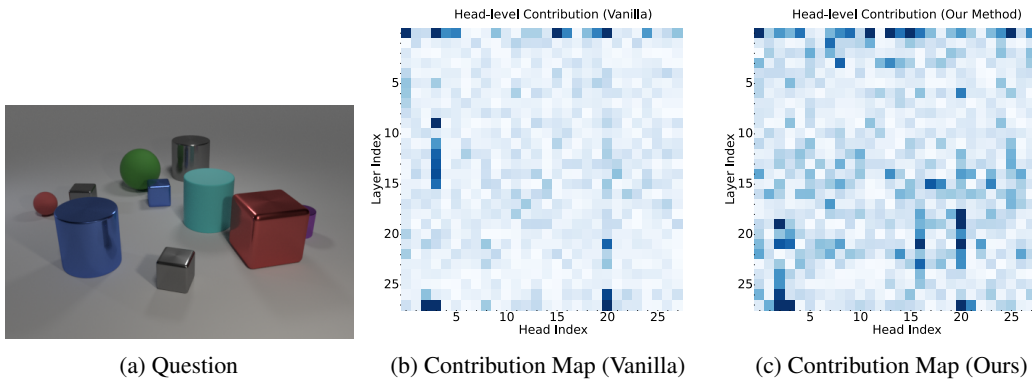
The image shows the square root of 2 as approximately **1.41421**. The image also shows various mathematical constants and numbers sequences, including the first 150 decimal places, and it provides many of the first decimal places as 1.4 ...

</think>

Answer: No. It's about 1.414, which is slightly above 1.41.

Analysis. In this case, we can clearly observe the differences in perception and reasoning abilities between the two methods. The Vanilla model, when tasked with identifying the value of the square root of 2, incorrectly perceived it as 1.44 rather than the correct value of 1.414. This perceptual error indicates insufficient contribution from the shallow perception heads, leading to poor sensitivity to numerical details. Furthermore, the subsequent reasoning process in the Vanilla model was unclear, with its step-by-step logic lacking coherence, and the final answer deviated from the original question. This combination of perceptual inaccuracy and weak reasoning aligns with the contribution map, which shows limited activation in both shallow perception heads and mid-to-deep reasoning heads. In contrast, our method demonstrates significant enhancements in both shallow perception heads and mid-to-deep reasoning heads, resulting in a more balanced improvement across perception and reasoning capabilities. Specifically, in this example, our method not only correctly identified the square root of 2 as 1.41421 but also precisely reasoned that the value is slightly above 1.41, thereby providing an accurate and contextually relevant answer. This highlights the superiority of our approach in achieving **improvements in both perceptual accuracy and logical reasoning**.

H.5 CASE IV: COMPREHENSIVE HALLUCINATION



Question - Image (a)

There is a big thing that is the same color as the tiny ball; what shape is it?

Vanilla - Image (b)

<think>

Okay, so I'm trying to solve this problem where I need to figure out the shape of a big object that's the same color as a tiny ball. Let's start with understanding the question carefully. The word "same color" probably refers to the same color group, meaning, or category, but I think it might be a bit too complicated.

The image shows a 3D geometric composition with a **large red sphere**, which is the "same color" of the small red sphere. Hmm, so the question is asking for the shape of a large object that's the same color as the tiny ball. So the tiny ball is a sphere, and the large one the same color as the tiny ball would also be **a sphere as well**, but much bigger in size.

The image shows a **large red sphere** within other shapes, so the **large red sphere** stands out the same color of another 3D object. The only other 3D object that could be described as the same color would also be **another sphere**. The image shows a **large red sphere**, where another 3D object can be a sphere. Therefore, the **large red sphere** stands out the same color of another sphere, which would also be a sphere as well. So, the **large red sphere** is a sphere, and another 3D object sharing the same color can be **a sphere** as well.

So the answer should be a sphere.

</think>

The image shows a **large red sphere**, which is the same color of another sphere. Therefore, the **large red sphere** is a sphere, and the one same color would be a sphere as well. The image shows a **large red sphere**, with another shape being a **large red sphere**. So, the **large red sphere** is a sphere, and another 3D object can be a sphere as well.

****Answer:** Sphere**

Ours - Image (c)

<think>

Okay, so I'm trying to figure out the answer to this question: There is a big thing that is the same color as the tiny ball; what shape is it?

The image shows a big red cube next to a tiny red sphere. The tiny ball is a sphere, and the big thing mentioned is a red cube. Since the image shows a big red cube next to a tiny red sphere, the big thing must be a red cube.

Therefore, the answer is a red cube.

</think>

The image shows a big red cube next to a tiny red sphere. The tiny ball is a sphere, and the big thing mentioned is a red cube.

Answer: The shape is a red **cube**.

Analysis. The Vanilla model’s reasoning exhibits systemic failure: it is convoluted, opaque, and internally inconsistent. Contribution maps reveal diffuse attention without stable anchoring, producing weak and unreliable signals across both perceptual and relational heads. More critically, attribute cross-talk collapses the distinction between color and shape, preventing the model from executing the necessary reasoning pipeline of anchoring → same-color filtering → applying the “big” constraint. As a result, it defaults to a spurious shortcut and outputs the incorrect label “sphere.” In contrast, our method imposes structured coordination: shallow heads focus sharply on color and size cues, while deep heads strengthen cross-object reasoning that jointly enforces the predicates “same color as the tiny ball” and “big,” leading to the correct identification of the large red cube. These results highlight that without principled alignment between perception and reasoning, models are prone to brittle and misleading inference.

REFERENCES

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

-
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.
- Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoub, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2023.
- Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.
- Lingfeng Ming, Yadong Li, Song Chen, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Ocean-r1: An open and generalizable large vision-language model enhanced by reinforcement learning. <https://github.com/VLM-RL/Ocean-R1>, 2025. Accessed: 2025-04-03.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.
- Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- Ke Wang, Juntong Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.
- Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.