CVIF: A CRITICALITY-DRIVEN VISUAL INTERVENTION FRAMEWORK FOR GEOMETRIC DIAGRAM UNDERSTANDING IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite significant progress in visual tasks by Multimodal Large Language Models (MLLMs), geometric diagram understanding remains challenging due to the presence of sparse visual cues and ambiguous symbol-primitive associations. MLLMs may therefore rely on textual priors, producing interpretations that conflict with visual evidence. We introduce the training-free Criticality-Driven Visual Intervention Framework (CVIF), an inference-time method that localizes critical layers and executes visual interventions during the transition from evidence aggregation to semantic decoding. At these layers, a Geometry-Constrained Local Relation Reconstruction (GCLR) module selects and weights vertex-centered visual evidence, while an Adaptive Visual Steering Operator (AVSO) redistributes attention mass toward the selected tokens. Experiments on PGPS9K and PGDP5K show that CVIF raises Overall F1 from 77.85 to 85.58 and from 75.23 to 82.84, respectively, establishing a novel inference-time visual intervention paradigm.

1 INTRODUCTION

In recent years, Multimodal Large Language Models (MLLMs) have achieved substantial progress in the field of multimodal understanding (Lu et al., 2024; Wang et al., 2024; Zhang et al., 2025; Qiao et al., 2025; Chen et al., 2026). Geometric Problem Solving (GPS), which integrates geometric diagrams with textual questions, has emerged as a crucial benchmark for evaluating their cognitive capabilities (Ma et al., 2026; Chen et al., 2021). However, existing MLLMs tend to deduce visual information for geometric reasoning from textual contexts rather than genuinely perceiving the diagrams. Extensive empirical studies reveal that the current end-to-end GPS paradigm (Zhang et al., 2024a;b) conflates perception errors with reasoning failures, making it difficult to isolate the true source of model limitations (Sun et al., 2026; Zhang et al., 2024b; Guo et al., 2025; Wang et al., 2026; Cho et al., 2025). Multiple diagnostic studies consistently confirm that perception errors remain the primary cause of failures in geometric reasoning tasks (Wang et al., 2025a;b) (Figure 1(a)). Therefore, it is imperative to explicitly decouple visual perception from logical reasoning by requiring models to independently parse geometric diagrams and generate formal language descriptions, thereby enabling an objective and unambiguous evaluation.

Why do MLLMs readily make errors in perceiving geometric diagrams? Our study finds that these errors can be traced to: (1) visual cues captured at early layers being overlooked in deeper layers, and (2) attention mechanisms in deeper layers becoming excessively focused on letter symbols while overlooking geometric structures. This is precisely consistent with recent observations from extensive research on MLLM vision that such failures are primarily not caused by an inability to perceive visual information (Khayatkhoei et al., 2025; Liu et al., 2026; Shao et al., 2026). First, attending to relevant visual tokens does not ensure their continued use in later layers. As illustrated in Figure 1(b), textual priors can favor an incorrect interpretation of sparse visual inputs (Favero et al., 2024; Liu et al., 2024). Subsequent attention updates may reinforce features consistent with that interpretation while reducing the contribution of conflicting visual evidence. We refer to this reinforcement of an existing interpretation as confirmation bias in the following discussion. Second, in geometric tasks, attending to the correct symbols does not by itself establish the geometric relations between them. As shown in Figure 1(c), recognizing a number or a vertex label is insufficient to determine which incident edges define the annotated angle. Human perception combines prior knowledge with

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.

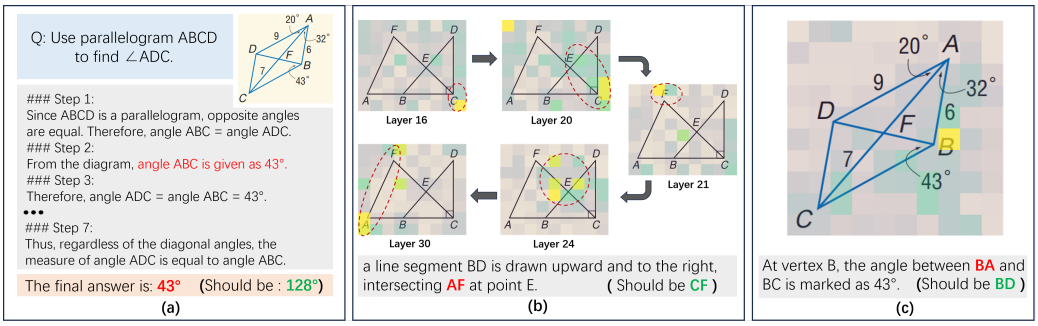

Figure 1: Attention failure modes of MLLMs in geometric diagram perception. (a) Perceptual errors during the process of geometric reasoning lead to prediction failures. (b) When predicting which line segment intersects BD at point E, shallow layers attend to the visually correct point C. As depth increases, attention shifts toward evidence that confirms a prior-compatible candidate (e.g., AF); by a later layer, the visually correct point C receives little attention. (c) When determining which angle is annotated by $43°$, late-layer attention may concentrate on the label "B" and nearby symbols while failing to recover the local topology of the three edges incident to B. In particular, the model does not sufficiently use edge BD as structural evidence for resolving the angle assignment.

incoming sensory evidence. When the two conflict, visual evidence can revise an initial expectation before a final judgment is formed. By contrast, an MLLM may increasingly favor a prior-consistent interpretation during autoregressive generation, even when the diagram contains conflicting visual cues. Once these cues lose influence in the model's internal representations, output-level correction alone cannot reliably restore them. This observation motivates intervention inside the model, rather than only through external prompting or post-hoc correction.

Building on the above analysis, we aim to design an active intervention mechanism that can identify the critical locations for visual intervention and provide relevant geometric cues. First, conceptualizing Transformer inference as an energy-descent process, we can view the failure illustrated in Figure 1 as a consequence of hidden states failing to sufficiently leverage visual cues to establish the direction of energy descent. As a result, hidden states are overly misled by textual priors and gradually become trapped in a local-minimum state. We introduce the concept of representational criticality to describe a state in which a hidden representation has integrated diverse cues but has not yet formed a stable preference for a specific interpretation. Representations in such a state are particularly sensitive to changes in the relative contribution weights of visual evidence and textual priors, making them ideal locations for visual intervention. Second, in geometric diagrams, geometric vertices not only serve as anchors for linguistic descriptions but also act as key nodes for constructing higher-level geometric topological structures. We can therefore use geometric vertices as entry points and combine them with the model's attention history to identify relevant visual evidence that encodes local topological information. Building on this idea, we propose CVIF, a training-free framework that uses layer-wise prediction diagnostics to select intervention layers and steers attention toward vertex-guided visual features. Our main contributions are: (i) We propose a Criticality-Driven Visual Intervention Framework, which mitigates confirmation bias and visual hallucinations in MLLMs by applying mechanism-level perturbations at the criticality. (ii) We design a Geometry-Constrained Local Relation Reconstruction (GCLR) mechanism, which reconstructs topological relations within the high-dimensional semantic space. (iii) We introduce an Adaptive Visual Steering Operator (AVSO), which safely reshapes multi-head attention distributions, enabling the model to resist deep-layer textual inertia without disrupting native logical reasoning.

2 Related Work

Geometric Problem Solving and Geometry Diagram Parsing. Geometric Problem Solving has long served as a core benchmark for evaluating the complex cognitive and spatial reasoning capabilities of artificial intelligence (Lu et al., 2021; Seo et al., 2014). Recently, MLLMs have demonstrated strong capabilities on various mathematical reasoning benchmarks (Lu et al., 2024; Qiao et al., 2025). However, their performance on GPS tasks remains unsatisfactory (Xu et al., 2025). Most

existing GPS works rely on end-to-end benchmarks, which tend to conflate perception errors with reasoning failures (Zhang et al., 2024a;b). In fact, recent studies have identified that MLLMs often suffer from severe perceptual impairments during long-sequence generation (Peng et al., 2026). For geometric diagrams, this challenge is further exacerbated by their extremely sparse and abstract topological features. Models may rely on textual priors when interpreting geometric diagrams, producing descriptions that are plausible from the text but inconsistent with the visual evidence. To address this, Geometry Diagram Parsing approaches attempt to translate diagrams into formal languages (Zhang et al., 2022; Wang et al., 2026) to decouple perception from reasoning.

Mechanistic Interpretability and Representation Dynamics. Mechanistic interpretability studies seek to characterize the computations implemented by neural networks(Elhage et al., 2021). Layer-wise analyses identify depth-dependent linguistic representations(Tenney et al., 2019) and show that feed-forward networks (FFNs) can function as key-value memories whose outputs contribute to vocabulary-space predictions(Geva et al., 2021; 2022). Lad et al. (2026) propose stages of inference involving contextual feature construction, prediction ensembling, and final distribution sharpening. These studies motivate a functional interpretation of changes across layers, rather than a uniform view of the forward pass. Connections between attention and modern Hopfield networks also provide an energy-based perspective on associative retrieval(Ramsauer et al., 2021). In multimodal settings, studies report reliance on textual priors and limitations in visual grounding(Deng et al., 2025; Lee et al., 2025; Tong et al., 2024). Together, these findings motivate investigating where visual evidence can be reinforced during prediction formation.

3 Analysis of Intervention Locations and Geometric Cues

3.1 Critical Locations for Visual Intervention

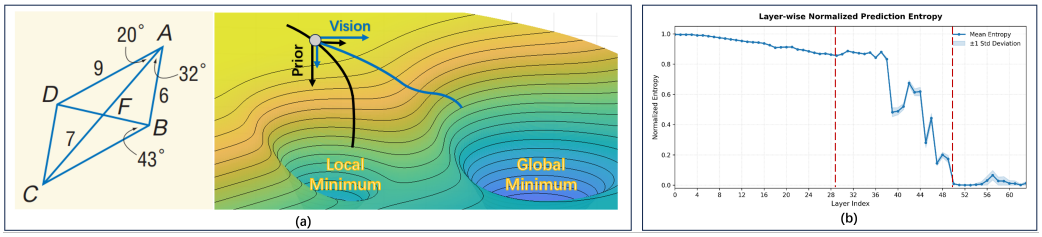

Figure 2: Prior-induced local minima and vision-guided trajectory correction. (a) When examining the angle at vertex B, textual priors tend to focus on the conventional $\angle$ ABC, whereas visual perception is naturally drawn to the salient feature "43°". As depicted in the energy landscape, this causes the prediction state to be pulled by prior forces (black arrow), making it highly susceptible to being trapped in a neighboring local minimum that represents an erroneous prediction (i.e., $\angle ABC = 43°$). During the critical state, introducing the local topological structure at vertex B provides a vision-driven steering force (blue arrow), thereby reducing the confidence of local minima induced by prior and steering the prediction toward the global minimum. (b) Layer-wise normalized predictive entropy for Qwen3-VL-32B at the question-end token.

To conceptualize the dynamics of forward inference, we adopt an energy-landscape perspective inspired by continuous Hopfield networks (Ramsauer et al., 2021). From this viewpoint, we treat the forward pass as a process in which hidden states continually evolve toward stable states in a high-dimensional semantic space. For geometric diagram understanding, interpretations supported by textual priors and those supported by local visual evidence may compete with each other (Figure 2(a)). At representational criticality, small changes in the relative contributions of visual evidence and textual priors can alter the subsequent evolution of hidden states. Representational criticality therefore identifies a critical location for visual intervention: if intervention occurs too early, the model has not yet formed a semantic judgment relevant to the current prediction; if it occurs too late, the prior-favored interpretation has already become stable, making it difficult for local visual evidence to redirect the prediction trajectory. To identify these critical locations, we quantitatively track changes in hidden-state predictions across network depth. Let the hidden state at the l-th Transformer layer be $\mathbf{h}^{(l)} \in \mathbb{R}^d$. Mapping it to the vocabulary V through the language model head

$\mathbf{W}_{\text{lm}}$ yields the predictive distribution $P^{(l)} = \text{Softmax}(\mathbf{W}_{\text{lm}}\mathbf{h}^{(l)})$. We use normalized predictive entropy to characterize the uncertainty of the prediction state at layer l:

$$\mathcal{H}^{(l)}_{\text{norm}} = -\frac{1}{\log|V|}\sum_{v\in V} P^{(l)}(v)\log P^{(l)}(v). \tag{1}$$

The entropy trajectory reflects three stages (Figure 2(b)). First, the model aggregates multimodal evidence in a high-entropy state. It then enters a stage of semantic decoding and prediction formation, during which the entropy curve deviates from its steady trend and exhibits pronounced nonlinear changes. Finally, the model consolidates and refines candidate predictions, causing entropy to decrease to a low level and gradually converge. Based on the geometric inflection points of the entropy curve, we use linear fitting to identify a set of critical layers $\mathcal{L}$ at which visual intervention is applied.

3.2 Vertex-Centered Geometric Cues

Unlike natural images, angles, intersections, and collinearity relations in geometric diagrams are typically expressed jointly by discrete vertices, thin line segments, and nearby annotations. Such local topological information is visually subtle, yet crucial for correctly parsing geometric relations. During feature aggregation, attention tends to focus more readily on letters and numbers with explicit linguistic semantics, while weak edges, intersections, and their relative positional relations are easily overlooked. Geometric vertices provide a natural organizing center for addressing this issue. Visually, edges, angles, intersections, and higher-order local configurations are organized around their incident vertices. Linguistically, geometric expressions such as BD and $\angle ABC$ also refer to diagram entities through vertex labels. A vertex relevant to the current prediction can therefore be regarded as the center of a local semantic cluster: the visual representations surrounding it jointly encode the vertex label, incident edges, nearby annotations, and their topological relations (Figure 3). Because a vertex token covers a local image region and has been contextualized by the vision encoder, it can encode information about nearby incident edges and local structures. Preserving the salience of these vertex-centered representations enables subsequent FFN computations to organize nearby geometric evidence around the relevant vertices, thereby helping the model recover geometric relations such as edge connectivity, angle structures, and multi-edge intersections.

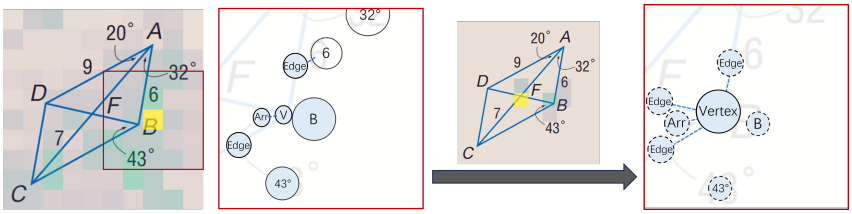

Figure 3: The Geometry-Constrained Local Relation Reconstruction (GCLR) mechanism. Using vertex B as a spatial anchor, GCLR identifies its incident edges (l_1, l_2, l_3) and nearby geometric annotations. The resulting target token set and weight matrix explicitly compensate for the relative topological relations of the local structures.

4 Method

4.1 Criticality-Driven Visual Intervention Framework

As illustrated in Figure 4, the CVIF workflow proceeds as follows. Given an image-text prompt, we identify a set of critical layers $\mathcal{L}$ based on the prediction entropy of the question-end token. For each layer l, let $\mathbf{A}^{(l)}_{\text{curr}} \in \mathbb{R}^{H\times L_{\text{seq}}}$ denote the multi-head attention matrix, where H is the number of heads and L_{seq} is the sequence length. We define $\bar{\mathbf{A}} \in \mathbb{R}^{L_{\text{seq}}}$ as the historical smoothed attention from the preceding n layers (averaged across head dimensions), and $\mathcal{P}_{\text{vertices}}$ as the geometric vertex coordinates obtained from a vertex detector (Jocher et al., 2023). If $l \notin \mathcal{L}$, the model retains its native feed-forward path. Conversely, if $l \in \mathcal{L}$, the GCLR module leverages $\bar{\mathbf{A}}$ and $\mathcal{P}_{\text{vertices}}$ to compute a target token set Ω_{tgt} and a topological weight matrix $\mathbf{W}_{\text{target}}$:

$$\Omega_{\text{tgt}}, \mathbf{W}_{\text{target}} = \text{GCLR}(\bar{\mathbf{A}}, \mathcal{P}_{\text{vertices}}). \tag{2}$$

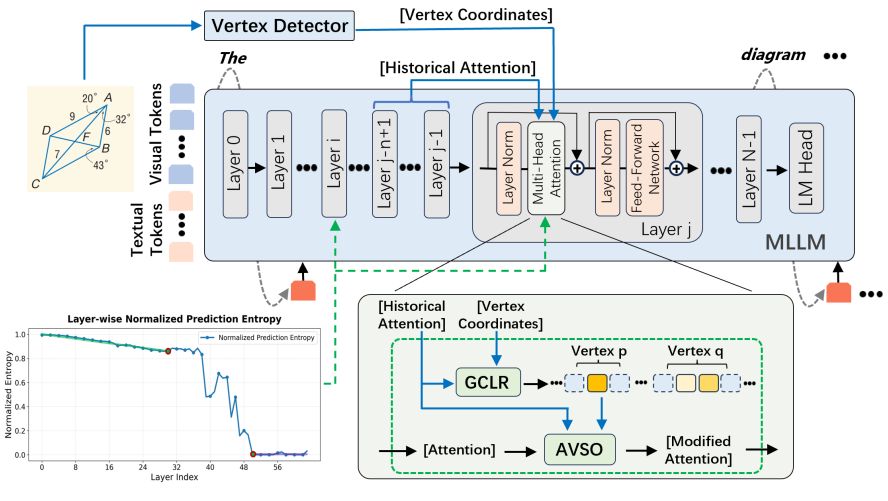

Figure 4: Overview of CVIF. Layer-wise normalized predictive entropy provides a diagnostic for selecting candidate critical layers during prediction formation. At these layers, GCLR utilizes geometric vertices as anchors to explicitly decouple and reconstruct spatial-semantic mappings for local structures. Subsequently, the AVSO performs distribution shaping and boost control over the attention, achieving active visual recall and prediction rectification.

AVSO then safely reshapes the original attention $\mathbf{A}^{(l)}_{\text{curr}}$ within the multi-head attention module:

$$\mathbf{A}^{(l)}_{\text{steered}} = \text{AVSO}(\mathbf{A}^{(l)}_{\text{curr}}, \bar{\mathbf{A}}, \Omega_{\text{tgt}}, \mathbf{W}_{\text{target}}), \tag{3}$$

The steered attention matrix $\mathbf{A}_{\text{steered}}$ is multiplied by the Value matrix, fusing key visual features to push the hidden state across the potential barrier in subsequent FFNs, steering subsequent representations toward visually supported predictions. Finally, to ensure rigorous quantitative evaluation, the generated geometric descriptions are translated into formal language using Qwen3-VL-32B.

4.2 Geometry-Constrained Local Relation Reconstruction

GCLR uses detected geometric vertices as anchors to select local visual evidence relevant to the current prediction. It first evaluates vertex relevance from historical attention and vertex-centered spatial priors, then identifies candidate tokens around selected vertices, and finally produces a target token set and topology-aware weights for subsequent attention steering.

Let $\mathcal{P}_{\text{vertices}} = \{\mathbf{p}_k\}_{k=1}^{N_v}$ denote the detected vertex coordinates, where N_v is the number of vertices. For each visual token $t \in \Omega_{\text{vis}}$, let $\mathbf{p}_t$ denote the center of its corresponding image region. The distance between vertex k and token t is defined as $d_k(t) = \|\mathbf{p}_t - \mathbf{p}_k\|_2$.

For each vertex k, we construct a topology prior and an anchor prior: $M^{(k)}_{\text{topo}}(t) = \rho \exp\left(-\frac{d_k(t)^2}{2\sigma_g^2}\right) + (1-\rho)\exp\left(-\frac{d_k(t)}{\sigma_l}\right)$, $M^{(k)}_{\text{anch}}(t) = \exp\left(-\frac{d_k(t)^2}{2\sigma_g^2}\right)$, where $\rho = 0.2$ controls the mixture of the two distance-decay components, and σ_g and σ_l control their spatial scales. The base relevance of vertex k is obtained through the weighted score of its historical attention under topological priors:

$$s^{(k)}_{\text{base}} = \sum_{t \in \Omega_{\text{vis}}} \bar{\mathbf{A}}(t)\, m^{(k)}_{\text{topo}}(t). \tag{4}$$

To incorporate distance-decay compensation, let d^{ref}_k denote the distance between vertex k and the prediction-relevant reference location. The final vertex score is

$$s^{(k)}_{\text{v}} = s^{(k)}_{\text{base}} + \eta \max_{u \in \{1,\dots,N_v\}} s^{(u)}_{\text{base}} \exp\left(-\frac{d^{\text{ref}}_k}{\tau}\right), \tag{5}$$

where η controls the compensation strength and τ controls its decay rate. We retain vertices whose scores exceed a dynamic threshold: $\mathcal{V}_{\text{act}} = \left\{k \,\middle|\, s^{(k)}_{\text{v}} \geq \mu_{\text{v}} + \alpha_{\text{v}}\sigma_{\text{v}}\right\}$, where μ_{v} and σ_{v} are the mean

and standard deviation of the candidate vertex scores, respectively. The normalized vertex weight is $\omega_{\text{v}}^{(k)} = \frac{s_{\text{v}}^{(k)}}{\sum_{u \in \mathcal{V}_{\text{act}}} s_{\text{v}}^{(u)}}, k \in \mathcal{V}_{\text{act}}$. For each activated vertex k, the token-level response is

$$s_{\text{tok}}^{(k)}(t) = M_{\text{topo}}^{(k)}(t)\bar{\mathbf{A}}(t), \qquad t \in \Omega_{\text{vis}}. \tag{6}$$

After applying the dynamic threshold $\mu_{\text{tok}}^{(k)} + \alpha_{\text{tok}}\sigma_{\text{tok}}^{(k)}$, we retain the representative tokens as $\mathcal{C}_{\text{feature}}^{(k)}$. To preserve the vertex-centered anchor, we additionally select $c_k = \arg\max_{t \in \Omega_{\text{vis}}} M_{\text{anch}}^{(k)}(t)\bar{\mathbf{A}}(t)$. The final candidate set for vertex k is $\mathcal{U}_k = \mathcal{C}_{\text{feature}}^{(k)} \cup \{c_k\}$. Within $\mathcal{U}_k$, we combine anchor localization and historical attention to compute token-level weights:

$$\omega_{\text{tok}}^{(k,j)} = (1-\gamma)\frac{m_{\text{anch}}^{(k)}(j)}{\sum_{u \in \mathcal{U}_k} m_{\text{anch}}^{(k)}(u)} + \gamma\frac{\bar{\mathbf{A}}(j)}{\sum_{u \in \mathcal{U}_k} \bar{\mathbf{A}}(u)}, \qquad j \in \mathcal{U}_k, \tag{7}$$

where γ balances spatial anchoring and historical attention. The final target token set is $\Omega_{\text{tgt}} = \bigcup_{k \in \mathcal{V}_{\text{act}}} \mathcal{U}_k$. For each target token $t \in \Omega_{\text{tgt}}$, we aggregate its contributions from all activated vertices:

$$\mathbf{W}_{\text{target}}(t) = \sum_{k \in \mathcal{V}_{\text{act}}} \omega_{\text{v}}^{(k)} \omega_{\text{tok}}^{(k,t)} \mathbb{I}[t \in \mathcal{U}_k]. \tag{8}$$

4.3 Adaptive Visual Steering Operator

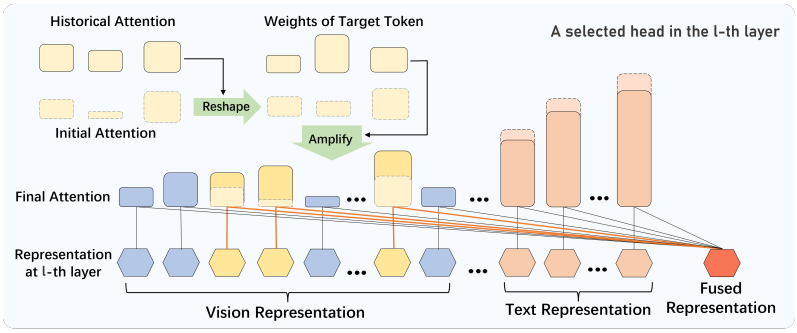

Figure 5: The principle of the Adaptive Visual Steering Operator (AVSO).

As illustrated in Figure 5, AVSO reshapes attention at critical layers without modifying model parameters. It uses the target-token set and topology-aware weights produced by GCLR, together with historical and current attention, to increase the influence of key geometric evidence while compensating for the added attention mass through text tokens.

For each target token $t \in \Omega_{\text{tgt}}$, $\mathbf{A}_{\text{curr}}^{(l)}(t)$ denotes its current attention values across heads, $\bar{\mathbf{A}}(t)$ denotes its head-aggregated historical attention prior, and $\mathbf{W}_{\text{target}}(t)$ denotes its normalized topology-aware weight. All vector operations below are applied element-wise along the head dimension. AVSO first compares current attention with the historical attention prior to compute a shaping factor for each target token:

$$\mathbf{M}_{\text{shape}}^{(l)}(t) = (1-\alpha)\mathbf{1} + \alpha\frac{\bar{\mathbf{A}}(t)}{\mathbf{A}_{\text{curr}}^{(l)}(t) + \epsilon}, \qquad t \in \Omega_{\text{tgt}}, \tag{9}$$

where $\alpha \in [0,1]$ controls the contribution of historical attention, ϵ is a numerical stability constant, and $\mathbf{1}$ is an all-one vector over attention heads. The shaped target attention is then defined as

$$\mathbf{A}_{\text{shaped}}^{(l)}(t) = \mathbf{A}_{\text{curr}}^{(l)}(t) \odot \mathbf{M}_{\text{shape}}^{(l)}(t), \qquad t \in \Omega_{\text{tgt}}, \tag{10}$$

where $\odot$ denotes element-wise multiplication. AVSO selectively triggers intervention only for heads exhibiting high attention toward visual tokens yet insufficient alignment with the target tokens.

$$\mathbf{\Phi}_{\text{gate}}^{(l)} = \mathbb{I}\left(\sum_{t \in \Omega_{\text{vis}}} \mathbf{A}_{\text{curr}}^{(l)}(t) \geq \tau_{\text{down}} \ \wedge \ \sum_{t \in \Omega_{\text{tgt}}} \mathbf{A}_{\text{curr}}^{(l)}(t) \leq \tau_{\text{up}}\right), \tag{11}$$

where τ_{down} and τ_{up} are the lower visual-attention threshold and upper target-attention threshold, respectively. To ensure that the shaped target attention reaches a competitive visual-attention scale, we compute a guarantee coefficient: $\boldsymbol{\gamma}^{(l)}_{\text{guarantee}} = \text{Clamp}\left(\frac{\max_{t\in\Omega_{\text{vis}}} \mathbf{A}^{(l)}_{\text{curr}}(t)}{\min_{t\in\Omega_{\text{tgt}}} \mathbf{A}^{(l)}_{\text{shaped}}(t)+\epsilon},\ 1.0,\ +\infty\right)$. For each target token, AVSO updates its attention as

$$\mathbf{A}^{(l)}_{\text{boosted}}(t) = \mathbf{A}^{(l)}_{\text{curr}}(t) \odot \left[\boldsymbol{\Phi}^{(l)}_{\text{gate}} \odot \mathbf{M}^{(l)}_{\text{shape}}(t) \odot \boldsymbol{\gamma}^{(l)}_{\text{guarantee}} \cdot \lambda \cdot \mathbf{W}_{\text{target}}(t) + \left(\mathbf{1} - \boldsymbol{\Phi}^{(l)}_{\text{gate}}\right)\right], t \in \Omega_{\text{tgt}}, \quad (12)$$

where $\lambda > 0$ controls the intervention strength. The attention mass added to target tokens is $\Delta\mathbf{P}^{(l)} = \sum_{t\in\Omega_{\text{tgt}}} \left(\mathbf{A}^{(l)}_{\text{boosted}}(t) - \mathbf{A}^{(l)}_{\text{curr}}(t)\right)$. Let the current attention mass assigned to text tokens be $\mathbf{m}^{(l)}_{\text{text}} = \sum_{t\in\Omega_{\text{text}}} \mathbf{A}^{(l)}_{\text{curr}}(t)$. The text-token scaling factor is defined as $\boldsymbol{\beta}^{(l)}_{\text{text}} = \text{Clamp}\left(\mathbf{1} - \frac{\Delta\mathbf{P}^{(l)}}{\mathbf{m}^{(l)}_{\text{text}}+\epsilon},\ 0,\ 1\right)$. The compensated text attention is

$$\mathbf{A}^{(l)}_{\text{text}}(t) = \mathbf{A}^{(l)}_{\text{curr}}(t) \odot \boldsymbol{\beta}^{(l)}_{\text{text}}, \qquad t \in \Omega_{\text{text}}. \quad (13)$$

Combining the boosted attention of target tokens, the compensated attention of text tokens, and the unchanged attention of all remaining tokens, followed by normalization over the token dimension, yields the final steered attention $\mathbf{A}^{(l)}_{\text{steered}}$.

5 Experiment

5.1 Experimental Setup

We evaluated the performance of geometric diagram parsing on 2,650 image–formal language pairs from PGPS9K (Zhang et al., 2023) and 2,650 pairs from PGDP5K (Hao et al., 2022). To enable unified evaluation, we convert PGDP5K annotations into the PGPS representation. By comparing the predicted set with the ground-truth set, we report precision (P), recall (R), and F1 for structural, semantic, and overall predicates. CVIF uses Qwen3-VL-32B-Instruct (Bai et al., 2025a) as the intervention backbone. We compare it with open-source models from the Qwen, InternVL, and GLM families (Bai et al., 2025b;a; Wang et al., 2025c; Team, 2026; Team et al., 2025), as well as proprietary models (Anthropic, 2025; OpenAI, 2025; Google, 2025).

Table 1: Performance comparison on PGPS9K.

Methods	Structure			Semantic			Overall		
	P	R	F1	P	R	F1	P	R	F1
Open-source MLLMs									
Qwen2.5-VL-7B-Instruct	73.58	38.00	50.12	38.32	43.82	40.89	47.19	40.70	43.70
InternVL3.5-8B-Instruct	66.14	43.70	52.63	43.42	51.55	47.14	51.44	47.38	49.33
InternVL3.5-30B-Instruct	70.03	58.89	63.98	57.39	56.12	56.75	62.32	57.61	59.87
InternVL3.5-8B	68.90	54.26	60.71	61.52	59.06	60.26	64.00	56.49	60.01
Qwen2.5-VL-32B-Instruct	70.15	53.01	60.39	63.33	58.9	61.03	65.49	55.74	60.22
Qwen3-VL-8B-Instruct	72.75	52.12	60.73	73.56	66.82	70.03	71.18	58.92	64.47
GLM-4.6V-Flash	77.25	52.08	62.22	77.39	62.68	69.26	76.42	56.98	65.29
Qwen3.5-9B	80.45	66.94	73.08	78.38	75.87	77.10	77.60	71.08	74.19
Qwen3-VL-32B-Instruct	82.91	76.16	79.39	79.63	74.67	77.07	80.37	75.48	77.85
Closed-source MLLMs									
Claude-4.5-Sonnet	81.47	69.25	74.87	78.17	68.93	73.26	79.04	69.10	73.74
GPT-5.2-1211	95.52	90.66	93.03	88.93	84.68	86.75	91.98	87.90	89.89
Gemini-3-Flash	98.84	91.14	94.84	86.73	84.14	85.41	92.91	87.90	90.33
CVIF (Ours)	90.73	88.41	89.56	83.28	81.95	82.61	85.72	85.43	85.58

5.2 Diagram Parsing Results

CVIF improves Overall F1 over Qwen3-VL-32B-Instruct from 77.85 to 85.58 on PGPS9K and from 75.23 to 82.84 on PGDP5K (Tables 1 and 2). On PGPS9K, structural and semantic F1 improve by 10.17 and 5.54 points, respectively; the corresponding gains on PGDP5K are 9.77 and 5.95 points. CVIF also consistently improves overall precision and recall: by 5.35 and 9.95 points on PGPS9K,

and by 4.84 and 10.15 points on PGDP5K. The larger recall gains indicate more complete recovery of geometric predicates, while the simultaneous precision gains show that this improved coverage does not introduce additional spurious relations. CVIF achieves the best Overall F1 among the evaluated open-source models on both datasets.

Table 2: Performance comparison on PGDP5K.

Methods	Structure			Semantic			Overall		
	P	R	F1	P	R	F1	P	R	F1
Qwen2.5-VL-32B-Instruct	70.15	53.01	60.39	63.33	58.9	61.03	65.49	55.74	60.22
Qwen3-VL-8B-Instruct	73.16	51.04	60.13	71.15	60.15	65.19	69.89	55.15	61.65
GLM-4.6V-Flash	76.95	53.18	62.89	71.27	56.73	63.18	72.04	54.78	62.24
Qwen3.5-9B	79.72	65.21	71.74	77.98	72.45	75.11	76.70	68.48	72.36
Qwen3-VL-32B-Instruct	81.28	74.58	77.79	77.43	69.81	73.43	78.26	72.43	75.23
GPT-5.2-1211	94.03	89.47	91.69	84.50	79.67	82.01	89.20	85.04	87.07
Gemini-3-Flash	97.15	93.54	95.31	77.62	75.67	76.63	88.30	85.47	86.86
CVIF (Ours)	88.76	86.38	87.56	80.84	77.97	79.38	83.10	82.58	82.84

Table 3: Sample Accuracy and Perfect Parsing Rate (PPR) on PGPS9K and PGDP5K.

Methods	PGPS9K					PGDP5K				
	Lines	Circles	Structures	Semantics	PPR	Lines	Circles	Structures	Semantics	PPR
Qwen2.5-VL-32B-Instruct	36.57	77.51	34.75	38.42	21.43	34.64	78.64	32.60	38.42	19.36
Qwen3-VL-8B-Instruct	42.79	77.51	38.08	50.75	25.06	42.45	77.85	37.40	50.00	24.00
GLM-4.6V-Flash	43.70	80.60	39.09	49.81	25.77	46.34	80.00	39.70	51.02	25.40
Qwen3.5-9B	56.00	85.43	50.64	61.89	37.36	55.55	82.60	48.30	63.47	36.34
Qwen3-VL-32B-Instruct	64.23	92.57	62.38	58.79	43.47	64.38	87.62	58.08	58.79	40.11
GPT-5.2-1211	84.42	91.36	79.13	73.55	62.04	85.81	86.42	75.66	71.40	58.26
Gemini-3-Flash	89.32	91.77	83.92	77.21	67.77	94.23	89.85	85.02	68.15	60.68
CVIF (Ours)	79.09	95.62	77.25	69.96	59.47	80.00	88.72	72.23	69.77	55.77

5.3 Diagram-level Exact Match Evaluation

Category-level F1 does not guarantee that an entire diagram is parsed correctly. We therefore report Sample Accuracy (SA), which requires all predicates within a category to match the ground truth, and Perfect Parsing Rate (PPR), which requires all predicates across the diagram to be correct. As shown in Table 3, CVIF achieves the best results among the evaluated open-source configurations on all metrics. Compared with Qwen3-VL-32B-Instruct, CVIF raises PPR from 43.47 to 59.47 on PGPS9K and from 40.11 to 55.77 on PGDP5K. These results show that CVIF improves not only individual predicate recovery, but also the global completeness and correctness of diagram parsing.

Table 4: Ablation results on PGPS9K.

Methods	Structure			Semantic			Overall		
	P	R	F1	P	R	F1	P	R	F1
Qwen3-VL-32B (Baseline)	82.91	76.16	79.39	79.63	74.67	77.07	80.37	75.48	77.85
Qwen3-VL-32B 2-stage	84.29	78.92	81.52	81.11	77.85	79.45	81.11	77.85	79.45
Uniform Intv.	86.00	80.13	82.96	79.66	77.03	78.32	81.07	78.70	79.86
w/o GCLR	86.46	81.21	83.75	81.49	78.79	80.12	82.51	80.10	81.28
w/o AVSO	86.12	81.11	83.54	80.34	78.54	79.43	81.12	79.92	80.52
CVIF (Ours)	90.73	88.41	89.56	83.28	81.95	82.61	85.72	85.43	85.58

5.4 Ablation Study

Table 4 shows that the two-stage pipeline, which generates a natural-language description before formalizing it, reaches 79.45 Overall F1, substantially below CVIF (85.58). Thus, CVIF's gains cannot be explained by changing the output pipeline alone. "Uniform Intv." replaces criticality-guided layer selection with fixed interventions at layers 16, 32, and 48 of the 64-layer backbone, yielding 79.86 Overall F1 versus 85.58 for CVIF and supporting criticality-guided placement. Replacing GCLR with Top-K visual-token selection reduces Overall F1 to 81.28, while replacing AVSO with direct attention amplification reduces it to 80.52. These results verify the importance of both criticality-guided layer selection and the two proposed modules.

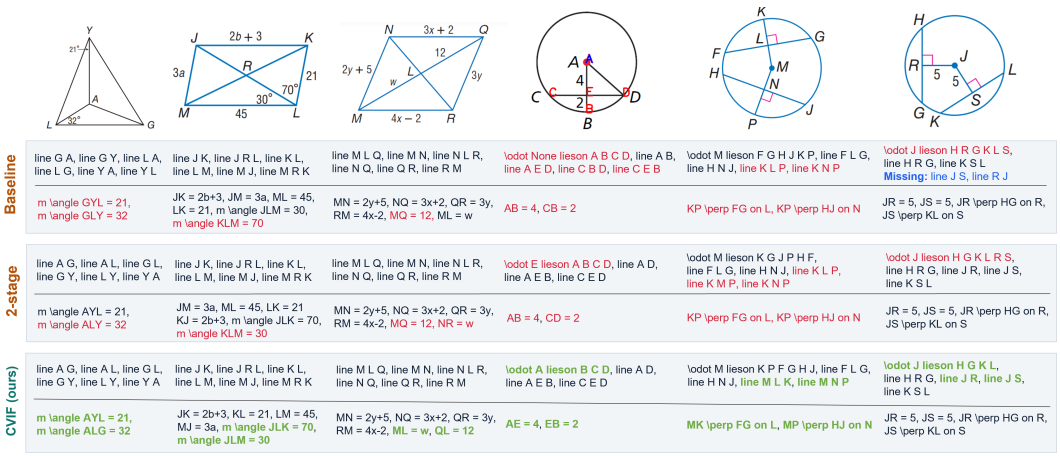

Figure 6: Qualitative case analysis on representative geometric diagrams.

5.5 Case Analysis

Figure 6 compares the parsing results of the baseline, the two-stage pipeline, and CVIF on representative diagrams. Although the baseline and two-stage pipeline recover many individual predicates, they often assign attributes to incorrect geometric entities or miss local structural relations required for correct parsing. In contrast, CVIF consistently recovers the relevant vertex-centered topology, including angle assignments, line connectivity, collinearity, and perpendicular relations. These examples illustrate that the proposed visual intervention helps the model retain and utilize local geometric evidence, thereby producing structurally consistent diagram parses.

5.6 Downstream Geometry Reasoning

We evaluate the downstream utility of diagram representations on the Vision Dominant split of the Plane Geometry subset of MathVerse (Zhang et al., 2024b). This representative and challenging geometry-reasoning setting requires models to rely primarily on visual evidence, making it well suited to evaluate whether improved diagram perception transfers to downstream reasoning. We compare five reasoning models under three settings: direct answering, augmentation with a formal-language representation, and augmentation with a natural-language diagram description (Table 5). Natural-language descriptions improve accuracy for all five models, with gains from 1.76 to 5.88 points. Formal-language representations improve four models, while the preferred representation varies by backbone. These results show that explicit diagram representations can improve downstream geometry reasoning in visually demanding settings.

Table 5: Downstream reasoning accuracy on MathVerse.

Methods	Qwen3-VL-32B	Qwen3-VL-8B	InternVL3.5-8B	Qwen2.5-VL-7B	Qwen3-VL-4B
Direct Answer	53.73	45.10	41.18	42.16	41.18
+ Formal Language	53.73	49.61	45.29	46.08	41.37
+ Text Description	58.63	48.24	47.06	46.47	42.94

6 Conclusion

In this paper, we proposed a training-free Criticality-Driven Visual Intervention Framework to mitigate textual-prior bias in MLLMs' geometric diagram understanding. An energy-landscape perspective motivates reinforcing local visual evidence at representational critical states, where competing interpretations have not yet become stable. The framework uses layer-wise prediction diagnostics to select candidate critical layers and directs attention toward vertex-guided visual evidence. We designed the GCLR, which uses geometric vertices as anchors to reconstruct "spatial-semantic" mappings, overcoming the failure of global attention to capture complex topologies. Furthermore, by introducing the AVSO, our framework achieved robust distribution reshaping. The reported experiments show improved diagram-parsing performance with the intervention, and evaluate the use of generated descriptions in downstream geometry reasoning, demonstrating a novel inference-time intervention paradigm for mitigating visual hallucinations.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language polishing and LaTeX formatting, including adaptation to the conference template. The authors reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies geometric diagram understanding and does not involve human participants, personal data, or high-risk decision-making. The proposed framework aims to improve the reliability of visual perception by strengthening the use of geometric evidence during inference.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 provide the methodological details of CVIF, including critical-layer identification, geometry-constrained local relation reconstruction, and adaptive visual steering. The corresponding formulas and theoretical analyses are presented alongside the method description. Section 5 reports the experimental backbone, datasets, evaluation metrics, comparison settings, and ablation configurations used to evaluate the proposed framework.

REFERENCES

Anthropic. The Claude 3 and 4.5 model family, 2025. `https://www.anthropic.com`.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 513–523, 2021.

Shuhang Chen, Hangjie Yuan, Yunqiu Xu, Pengwei Liu, Tao Feng, Jun Cen, Zeying Huang, and Yi Yang. Mathflow: Enhancing the perceptual flow of mllms for visual mathematical problems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 967–992, 2026.

Seunghyuk Cho, Zhenyue Qin, Yang Liu, Youngbin Choi, Seungbeom Lee, and Dongwoo Kim. Geodano: Geometric vlm with domain agnostic vision encoder. *arXiv preprint arXiv:2502.11360*, 2025.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3867–3876. IEEE, 2025.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Anna Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312. IEEE, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Mor Geva, Avi Caciularu, Kevin Wang, Guy Dar, and Jonathan Berant. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45, 2022.

Google. Gemini 3 Pro, 2025. https://deepmind.google/models/gemini/pro/.

Shiwei Guo, Sihang Jiang, Qianxi He, Yanghua Xiao, Jiaqing Liang, Bi Yude, Minggui He, Shimin Tao, and Li Zhang. Do large language models truly understand cross-cultural differences? *arXiv preprint arXiv:2512.07075*, 2025.

Yihan Hao, Mingliang Zhang, Fei Yin, and Linlin Huang. PGDP5K: A diagram parsing dataset for plane geometry problems. In *Proceedings of the 26th International Conference on Pattern Recognition (ICPR)*, pp. 1763–1769, 2022.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. https://github.com/ultralytics/ultralytics.

Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski, et al. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 68194–68213, 2025.

Vedang Lad, Jin Hwa Lee, Wes Gurnee, and Max Tegmark. Remarkable robustness of llms: Stages of inference? *Advances in Neural Information Processing Systems*, 38:130050–130083, 2026.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision (ECCV)*, pp. 125–140. Springer, 2024.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 72849–72876, 2026.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Lihua Xia, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and natural language. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

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 *International Conference on Learning Representations (ICLR)*, 2024.

Jianzhe Ma, Wenxuan Wang, and Qin Jin. A survey of deep learning for geometry problem solving. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39416–39448, 2026.

OpenAI. GPT-5 system card. Technical report, OpenAI, 2025. https://openai.com/index/gpt-5-system-card/.

Ruiying Peng, Xueyu Wu, Jing Lei, Lu Hou, Yuanzheng Ma, and Xiaohui Li. Deeper thought, weaker aim: Understanding and mitigating perceptual impairment during reasoning in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Runqi Qiao, Qiuna Tan, Guanting Dong, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 20023–20070, 2025.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Georgo K Sandner, et al. Hopfield networks is all you need. In *International Conference on Learning Representations (ICLR)*, 2021.

Min Joon Seo, Hannaneh Hajishirzi, Ali Farhadi, and Oren Etzioni. Diagram understanding in geometry questions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 2831–2838, 2014.

Jiahao Shao, Yuanbo Yang, Yiyi Liao, Yujun Shen, Ceyuan Yang, and Yinghao Xu. Thinking with tools, not with pixels: Tool calls as text scaffolds for visual reasoning. *arXiv preprint arXiv:2608.09682*, 2026.

Yanpeng Sun, Shan Zhang, Wei Tang, Aotian Chen, Piotr Koniusz, Kai Zou, Yuan Xue, and Anton Hengel. Math blind: Failures in diagram understanding undermine reasoning in mllms. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 124385–124422, 2026.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, et al. GLM-4.5v and GLM-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024.

Ke Wang, Junting 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.

Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19541–19551, 2025a.

Peijie Wang, Chao Yang, Zhong-Zhi Li, et al. Solidgeo: Measuring multimodal spatial math reasoning in solid geometry. *arXiv preprint arXiv:2505.21177*, 2025b.

Peijie Wang, Ming-Liang Zhang, Jun Cao, et al. Geoparsing: Diagram parsing for plane and solid geometry with a unified formal language. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29876–29903, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL 3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

Liangyu Xu, Yingxiu Zhao, Jingyun Wang, et al. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning. *arXiv preprint arXiv:2504.12597*, 2025.

Jiaxin Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, Cheng-Lin Liu, and Yashar Moshfeghi. Geoeval: Benchmark for evaluating llms and multimodal models on geometry problem-solving. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1258–1276, 2024a.

M. Zhang et al. PGPS9K: A large-scale dataset for geometric problem solving and theorem proving. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

Ming-Liang Zhang, Fei Yin, Yi-Han Hao, and Cheng-Lin Liu. Plane geometry diagram parsing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1636–1643, 2022.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, et al. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *European Conference on Computer Vision (ECCV)*, pp. 169–186, 2024b.

Shan Zhang, Aotian Chen, Yanpeng Sun, Jindong Gu, Yi-Yu Zheng, Piotr Koniusz, Kai Zou, Anton van den Hengel, and Yuan Xue. Open eyes, then reason: Fine-grained visual mathematical understanding in mllms. *arXiv preprint arXiv:2501.06430*, 2025.