TRACING PERCEPTUAL TRANSFER IN VISION-LANGUAGE MODELS VIA MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent progress in visual mathematical reasoning, perception remains a major bottleneck for Vision-Language Models (VLMs). However, it remains unclear where perceptual transfer occurs within the visual hierarchy of VLMs and how its representational and functional effects vary across layers. To investigate this question, we use model merging as a tool and apply it to the perception modules of VLMs, while keeping the language model frozen, allowing us to understand how perceptual transfer occurs across layers. Across multiple VLMs, this intervention yields gains on several visual mathematical reasoning benchmarks (MathVista, MathVerse, MathVision), providing evidence that perceptual capabilities can be transferred without modifying the LLM. To understand where these gains arise, we further analyze perceptual transfer across the visual hierarchy. Layer-wise probing shows that primitive geometric attributes become linearly accessible relatively early in the visual hierarchy, whereas relation-sensitive information becomes more prominent in intermediate layers. We further perform backward and forward layer replacement to examine the functional effects of perceptual transfer. Removing merged parameters from intermediate layers causes the largest performance degradation, while inserting them into the base model yields the largest performance gains in the same layers. Together, these complementary analyses reveal a non-uniform layer-wise organization of perceptual transfer, with relation-sensitive information and its strongest functional effects concentrated in middle layers. These findings provide a layer-wise account of how perceptual knowledge is represented and transferred within VLMs, offering insights for targeted perception-side adaptation.

1 INTRODUCTION

Vision-Language Models (VLMs) have achieved remarkable progress in visual mathematical reasoning, driven by techniques such as Chain-of-Thought (CoT), supervised fine-tuning (SFT) and reinforcement learning Shi et al. (2024); Zhang et al. (2025a); Huang et al. (2025); Chen et al. (2025a); Ong et al. (2025). Perception, however, is equally fundamental to visual mathematical reasoning, as it extracts structured mathematical evidence from visual and textual input Yang et al. (2026b). This evidence forms the basis for reasoning, meaning that even models with strong reasoning capabilities may still reach incorrect conclusions when the perceptual information is not correctly extracted. Indeed, recent studies increasingly identify perception as a major bottleneck in visual mathematical reasoning Wu et al. (2024); Cai et al. (2024); Gao et al. (2025); Zhang et al. (2025b); Qiao et al. (2025); Ma et al. (2025); Liu et al. (2025).

Despite growing efforts to improve VLM perception, our understanding of how these improvements arise internally remains limited. Prior work has attempted to improve visual understanding through specialized encoder training Zhang et al. (2025c), fine-grained alignment tuning Chen et al. (2026), multi-agent collaboration Xie et al. (2026), and inference-time refinement Jiang et al. (2026). However, these approaches treat perception as a black box, as they optimize or refine perceptual capabilities based on downstream performance without tracing how perceptual representations change inside the model.

Because these approaches assess perceptual improvements primarily through downstream performance, higher performance cannot reveal what perceptual knowledge has changed, where these

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.

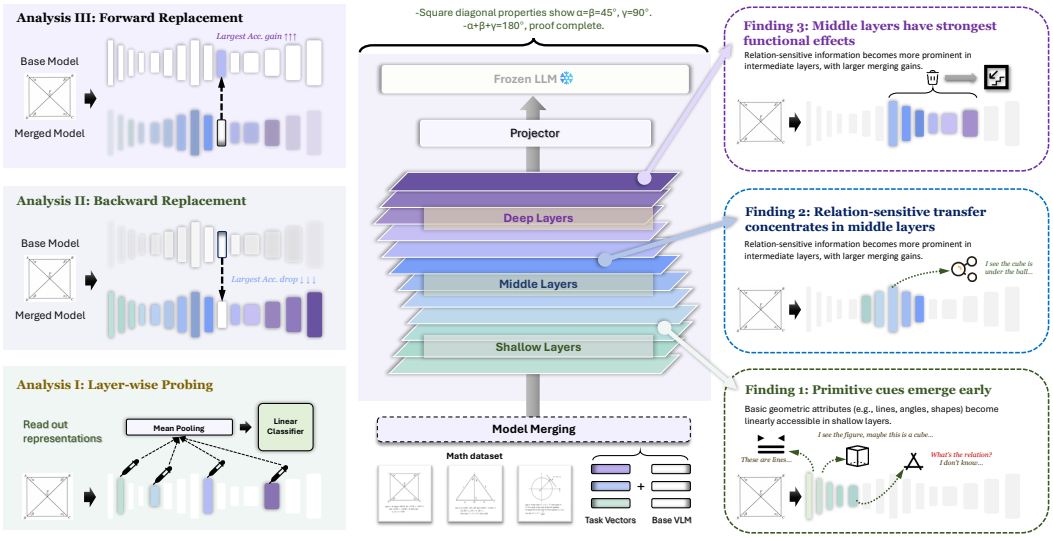


Figure 1: **Overview of perception-side transfer and layer-wise analysis.** We use model merging to introduce geometry-aware perceptual knowledge into the perception module while keeping the language model frozen, and analyze the resulting transfer through layer-wise probing and bidirectional layer replacement. The analyses reveal a non-uniform layer-wise pattern: primitive geometric cues are more linearly decodable in earlier layers, while relation-sensitive information and larger downstream effects are more prominent in intermediate layers.

changes are represented, or which layers mediate their effects on downstream tasks. This gap makes it difficult to diagnose the source of perceptual improvements and to determine where future interventions should target. Understanding the internal organization of perceptual knowledge is therefore important not only for explaining existing gains, but also for developing more targeted approaches to improving VLM perception.

To understand how perceptual knowledge is transferred across VLM layers, we require an intervention that can selectively modify the perception module which keeping the language model frozen. Such an intervention allows us to isolate perception-side changes and analyze their layer-wise effects. Model merging Ilharco et al. (2023) provides a natural solution to this requirement. By combining task vectors in parameter space, model merging enables targeted knowledge transfer without additional optimization during the merging process. Therefore, we use model merging Ilharco et al. (2023) as a lightweight, training-free intervention. Instead of modifying the language model, we inject geometry-aware task vectors into the perception modules of VLMs, which allows us to examine perception-side transfer separately from LLM updates. Specifically, we construct task vectors Ilharco et al. (2023) from perception modules tuned on diverse geometry datasets, including PGPS9K Zhang et al. (2023), Geo3K Lu et al. (2021), UniGeo Chen et al. (2022), and PGDP5K Hao et al. (2022), and merge them back into the original VLMs through simple linear arithmetic.

With this controlled perception-side intervention, we further analyze how perceptual transfer is organized across the visual layers. We consider two views. Layer-wise probing examines how geometric attributes are encoded across middle representations, while forward layer replacement and backward layer replacement measure how transferred parameters at different layers affect downstream performance. Across models, these analyses reveal a clear non-uniform layer-wise pattern. Primitive-level attributes become linearly decodable in earlier layers, whereas relation-level information becomes more prominent in middle layers. We further find that merging induces only limited changes in primitive-level attributes but more noticeable improvements in relation-level attributes. Consistently, replacing middle layers tends to produce larger changes in downstream performance than replacing shallow or deep layers. Together, these observations suggest that perceptual transfer is more closely associated with visual-layer representations that organize visual elements into geometric relations, rather than primarily altering low-level visual features.

Interestingly, although model merging is used primarily as an analysis intervention, we also observe improvements across several visual mathematical benchmarks, including gains on vision-dependent and geometry-sensitive subsets. Overall, our results provide a layer-wise view of how perceptual knowledge is represented and transferred within VLMs, and highlight the non-uniform roles of different visual layers in perception-side adaptation.

2 Related Work

2.1 Mathematical Reasoning in VLMs

Recent research has focused on enhancing the reasoning ability of VLMs for visual mathematical tasks Lewkowycz et al. (2022); Luo et al. (2025); Yue et al. (2024); Azerbayev et al. (2023). To achieve this, instruction-tuning and reinforcement learning (RL) frameworks have been widely deployed to improve the performance of VLMs in solving visual mathematical problems Zhang et al. (2025a); Huang et al. (2025); Luo et al. (2026); Chen et al. (2025a); Ong et al. (2025). However, a growing consensus has emerged that a major bottleneck of visual mathematical reasoning is perception Chen et al. (2021); Trinh et al. (2024). This has motivated a range of approaches for improving visual understanding Kahou et al. (2018); Kafle et al. (2018); Methani et al. (2020); Masry et al. (2022), including specialized encoder training Zhang et al. (2025c), fine-grained alignment tuning Chen et al. (2026), multi-agent collaboration Xie et al. (2026), and inference-time refinement Jiang et al. (2026). Our work follows this direction by using model merging as a lightweight intervention to study geometry-aware perceptual transfer and its layer-wise organization in VLMs.

2.2 Model Merging

Model merging combines parameters from pretrained or fine-tuned models to compose capabilities without additional optimization during the merging stage Yang et al. (2026a); Dong et al. (2026); Ma et al. (2026). Task Arithmetic Ilharco et al. (2023) represents fine-tuning-induced parameter changes as task vectors and provides a simple basis for linear capability composition. Subsequent methods, including TIES-Merging Yadav et al. (2023) and DARE Yu et al. (2024), improve merging by reducing parameter conflicts or sparsifying task vectors, while more recent approaches explore structured parameter representations such as task singular vectors Gargiulo et al. (2025) and difference vectors Wang et al. (2026).

Beyond capability composition, model merging has also begun to serve as a tool for studying how capabilities are organized and transferred within models. In VLMs, Chen et al. (2025b) use cross-modal merging to inject reasoning capabilities into the language component and analyze the layer-wise organization of perception and reasoning. Our work takes a perception-centered view: we keep the language model fixed and intervene only on the perception modules, using merging to induce controlled perceptual transfer. We then analyze how the transferred knowledge is represented and how its effects vary across visual layers through probing and layer replacement. For this purpose, we adopt linear task arithmetic as a simple and controlled merging operation rather than introducing a new merging algorithm.

3 METHOD – TRACING PERCEPTUAL TRANSFER

Our purpose is to understand how perceptual knowledge is transferred across VLM perception layers. First, we use perception-side model merging as an intervention and analyze the resulting transfer from two perspectives. Layer-wise probing measures the decodability of information in middle layers, while layer replacement evaluates the impact of transferred parameters on downstream performance.

3.1 Perception-side Model Merging

A Vision-Language Model (VLM) consists of three core components: a vision encoder, a projector for cross-modal alignment, and a large language model. While the vision encoder is responsible for extracting visual features to provide perceptual capabilities, the language model serves as the reasoning backbone for information processing and response generation. In this study, unlike prior

merging works that mainly inject reasoning abilities into the language model, we focus exclusively on the perception modules (encompassing both the vision encoder and the projection layer), whereas the language model remains frozen throughout the merging process. This design allows us to focus on perception-side parameter changes while reducing confounding from the language model.

To equip VLMs with specialized perceptual abilities while aiming to retain the original VLM's cross-modal alignment, we utilize the task arithmetic framework. First, let θ_{base} denote the parameters of the original pretrained perception modules. We define the multimodal alignment task vector τ_{vlm}, which captures the domain-general visual representation and alignment knowledge acquired by the VLM's perception modules during its instruction fine-tuning:

$$\tau_{\text{vlm}} = \theta_{\text{vlm}} - \theta_{\text{base}} \tag{1}$$

where θ_{vlm} represents the perception module parameters of the instruction-tuned VLM.

Similarly, let θ_{per} denote the parameters of the perception-tuned visual module. We define its corresponding perception task vector τ_{per}, which encapsulates the geometry-aware features learned during the perception tuning phase:

$$\tau_{\text{per}} = \theta_{\text{per}} - \theta_{\text{base}} \tag{2}$$

Rather than relying entirely on the VLM's original perception modules, we introduce the perception task vector τ_{per} to mitigate the geometric perception bottleneck. We integrate these geometry-aware capabilities into the VLM by linearly merging the two task vectors in the parameter space. The merged perception module θ'_{vlm} is formulated as:

$$\theta'_{\text{vlm}} = \theta_{\text{base}} + \lambda\tau_{\text{vlm}} + (1-\lambda)\tau_{\text{per}} \tag{3}$$

Here, $\lambda \in [0, 1]$ determines the weight assigned to the VLM multimodal alignment task vector and controls the trade-off between retaining the original VLM perception module and injecting geometry-aware perceptual features.

3.2 Layer-wise Transfer Analysis

To understand how geometry-aware perceptual transfer is distributed across the visual layers, we perform a layer-wise analysis from two perspectives. First, we examine the representational accessibility of geometric information by measuring whether task-relevant attributes can be linearly decoded from different visual layers. Second, we investigate the functional contribution of transferred parameters through layer-level replacement. Together, these analyses characterize perceptual transfer from both a representation perspective and a functional perspective.

Representation Decodability via Layer-wise Probing. To identify where perceptual knowledge becomes accessible within the visual layers, we introduce a layer-wise probing analysis. Directly attributing perception to a specific layer is challenging in visual mathematical reasoning, where perception and reasoning are highly entangled.

To obtain a controlled measurement of geometric representation decodability, we construct a synthetic geometry probing dataset with attribute prediction tasks, including line presence, angle type, line relation, and intersection count. Each sample consists of a rendered geometry diagram paired with its corresponding attribute label. The probing dataset is used exclusively for analysis and remains disjoint from both task vector construction and downstream benchmark evaluation.

For each visual layer, we mean-pool spatial tokens into a global feature representation and train an independent linear classifier for each attribute. The held-out classification accuracy measures whether the corresponding attribute is linearly decoded from that layer. We interpret higher probing accuracy as stronger decodability of information at a given layer, rather than direct evidence that the layer causally forms the corresponding capability. More details are provided in Appendix B.

Functional Impact via Layer Replacement. While layer-wise probing reveals where geometric information becomes accessible within the visual layers, it does not reveal how changes in transferred parameters at different layers affect downstream task performance.

To further understand the effects of perceptual transfer, we design two layer replacement experiments: backward replacement and forward replacement. For backward replacement, we replace each visual layer of the merged model with the corresponding layer from the base model and measure the resulting performance change. A performance decrease indicates that the removed layer contains transferred parameters that are important for the downstream task. For forward replacement, we replace each visual layer of the base model with the corresponding layer from the merged model and measure the resulting performance change. A performance increase indicates that the introduced transferred parameters at that layer contribute to the downstream task.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Base models and Task Vector Construction. In this paper, we employ LLaVA-Next-LLaMa3-8B (denoted as **LLaVA-8B**) Liu et al. (2024), InternVL3-8B (denoted as **InternVL-8B**) Zhu et al. (2025) and Qwen2.5-VL-7B (denoted as **Qwen-VL-7B**) Bai et al. (2025) as base models. For geometry datasets including PGPS9K Zhang et al. (2023), Geometry3K (denoted **Geo3K**) Lu et al. (2021), UniGeo Chen et al. (2022), and PGDP5K Hao et al. (2022), we build dataset-specific perception-tuned models from the corresponding base models. Each training sample is represented as a geometry diagram paired with a natural-language description of visible geometric primitives, symbols, and relations. Following the diagram-to-description annotation style introduced by GOLD Zhang & Moshfeghi (2024), we use the description only as a training target rather than an additional inference-time input. This training formulation enables the model to learn geometry-aware visual representations without direct supervision on complete problem solving or symbolic derivation. We emphasize that this perception-side tuning is used to construct task vectors. The subsequent merging operation is post-hoc and does not require gradient updates. If compatible geometry-specialized perception modules are available, this tuning stage can be skipped.

To perform perception-side model merging, we first construct task vectors from perception modules specialized for geometric understanding. Existing VLM checkpoints generally provide instruction-tuned models but do not include task-specific perception vectors required for our analysis. Therefore, we perform perception-side tuning on geometry datasets to obtain these task vectors. We emphasize that this tuning stage is only used for task vector construction. After obtaining the task vectors, the merging process itself is training-free and requires no additional gradient updates. Instead, it directly combines parameters in the perception module through task arithmetic, allowing us to study perception-side knowledge transfer under a controlled intervention.

Datasets. For training, we use four geometry datasets: PGPS9K Zhang et al. (2023), Geo3K Lu et al. (2021), UniGeo Chen et al. (2022), and PGDP5K Hao et al. (2022). These datasets cover diverse geometry diagrams, symbolic annotations, and problem types. To evaluate our method effectively, we evaluate on five benchmarks: MathVista (test-mini set)Lu et al. (2024), MathVerse Zhang et al. (2024), MMStar Chen et al. (2024), Dynamath Zou et al. (2025) and MathVision Wang et al. (2024). Additionally, using normalized problem text matching, option matching, and perceptual image hashing. We note that sample overlap exists between the training set and evaluation benchmarks. We exclude overlapped samples from the training set to prevent data leakage.

Implementation Details. All fine-tuned models are trained in BFloat16 precision using the AdamW optimizer. During the tuning phase, we set a learning rate of 2×10^{-4}, use a per-device batch size of 1 with gradient accumulation of 16, and train for 4 epochs on 10 NVIDIA RTX 3090 GPUs. We adopt a cosine learning-rate schedule with a warmup period covering 5% of the total training steps. During fine-tuning, only the vision-side parameters (vision encoder and projector) are updated during this process, while the language model remains frozen. For the linear merging of task vectors, we set the scaling factor λ to 0.8.

4.2 DOWNSTREAM PERFORMANCE

We evaluate the merged models across five benchmarks, with results summarized in Table 4.1. As shown in Table 4.1, for LLaVA-8B, merging generally improves mathematical reasoning perfor-

Model	Method	MathVista	MathVerse Benchmarks						MMStar		DM	MV
			Overall	T-D	T-L	V-I	V-D	V-O	All	Math		
LLaVA-8B	Baseline	37.4	20.1	25.9	20.8	21.1	16.5	16.0	**43.8**	30.0	22.7	13.8
	+ Geo3K	37.6	22.6	**29.1**	23.1	23.5	18.2	16.7	43.1	32.2	24.3	14.9
	+ UniGeo	37.5	21.4	26.5	21.4	23.0	17.1	15.8	42.9	31.0	23.5	14.1
	+ PGDP5K	37.2	22.1	28.5	22.5	22.0	16.8	15.5	42.1	31.8	23.8	14.5
	+ PGPS9K	**38.5**	**23.5**	27.8	**23.9**	**25.4**	**21.2**	**18.2**	42.4	**33.2**	**24.5**	**15.2**
InternVL-8B	Baseline	65.6	43.1	54.1	47.5	44.8	43.8	25.3	**67.3**	75.2	38.7	23.7
	+ Geo3K	65.4	43.8	**57.2**	48.2	45.1	42.5	24.1	65.5	76.5	39.2	24.0
	+ UniGeo	**66.5**	**47.6**	55.6	**51.2**	**50.1**	**51.5**	**28.7**	65.2	**78.6**	**40.8**	**26.3**
	+ PGDP5K	65.8	44.5	53.6	48.8	46.5	45.0	25.0	66.1	77.1	39.5	24.8
	+ PGPS9K	**66.5**	46.8	55.8	50.5	48.8	48.2	27.4	66.0	78.4	40.5	26.1
Qwen-VL-7B	Baseline	65.5	35.2	39.5	34.8	34.9	36.0	30.8	**64.2**	63.5	38.6	25.4
	+ Geo3K	65.4	36.8	**41.5**	36.1	35.4	36.5	31.0	63.2	64.8	39.7	26.1
	+ UniGeo	65.1	36.1	39.2	35.5	35.8	36.8	31.5	63.5	64.2	39.1	25.9
	+ PGDP5K	65.8	37.4	39.8	36.8	**38.2**	38.5	32.6	63.8	65.3	40.4	26.8
	+ PGPS9K	**66.3**	**37.8**	40.8	**36.9**	37.5	**39.4**	**33.1**	63.6	**65.4**	**40.5**	**26.9**

Table 1: The performance of LLaVA-Next-LLaMA3-8B model, InternVL3-8B model and Qwen2.5-VL-7B model with merged task vector across five benchmarks: MathVista (test-mini set), MathVerse (Overall, Text-Dominant, Text-Lite, Vision-Integrate, Vision-Dominant, and Vision-Only categories). MMStar (All and Math split), DynaMath (annotated as DM) and MathVision (annotated as MV). We **bold** the best result in each benchmark for each model. All results are averaged over three runs.

mance, with notable gains on geometry-sensitive, vision-dominant subsets. Specifically, the Geo3K-merged and UniGeo-merged variants improve the overall MathVerse score by 2.5% and 1.3%, respectively, while also improving DynaMath (DM) and MathVision (MV). The PGDP5K-merged model shows a similar trend, yielding a 2.0% gain on MathVerse overall and remaining competitive across multiple benchmarks. Among these variants, PGPS9K-merged LLaVA-8B achieves the strongest overall gains, including the best scores on MathVista, DM, and MV. It also improves the V-I, V-D, and V-O subsets of MathVerse by 4.3%, 4.7%, and 2.2%, respectively, indicating strong performance on visually grounded reasoning subsets.

Across backbones, we observe a similar but non-uniform pattern on InternVL-8B and Qwen2.5-VL-7B. Perception-side merging improves several visual mathematical benchmarks, but the gains vary across models and transfer sources. For example, UniGeo improves the V-D and V-O subsets of MathVerse by 7.7% and 3.4% on InternVL-8B, whereas PGPS9K improves the same subsets by 3.4% and 2.3% on Qwen2.5-VL-7B. Notably, the most effective transfer source also differs across backbones: UniGeo yields larger gains on several metrics for InternVL-8B, while PGPS9K is more effective for Qwen2.5-VL-7B. This variation suggests that the effect of perceptual transfer depends on the interaction between the transferred knowledge and the representations already present in the VLM.

Overall, these results show that perception-side transfer produces measurable but selective downstream performance gains rather than a uniform performance gain. The effects vary across tasks, backbones, and transfer sources, with larger improvements often appearing on vision-dependent subsets. However, benchmark scores alone cannot explain where these differences arise within the perception module. We therefore next examine the layer-wise organization of perceptual transfer.

4.3 Mechanism Analysis

Locating geometric information across layers. We use layer-wise probing to examine how geometric information is distributed across layers. As shown in Figure 2, we observe a consistent trend across Qwen-VL-7B and LLaVA-8B. Primitive-level attributes increase rapidly in early layers, peak around shallow-to-middle layers, and then remain stable or gradually decline. In contrast,

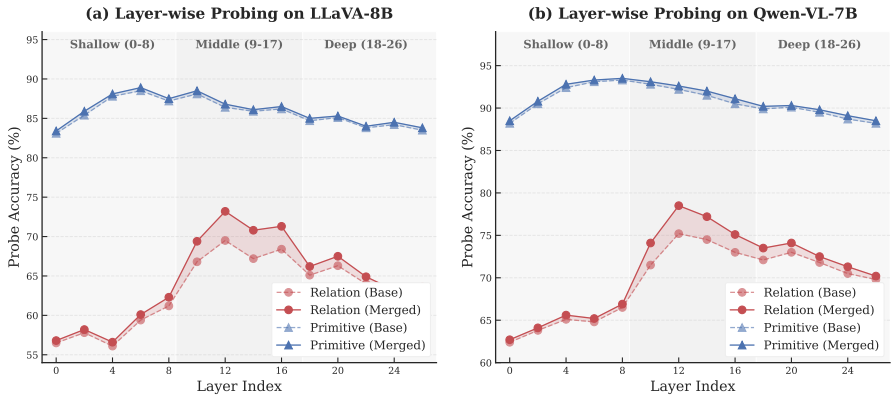


Figure 2: Layer-wise probing accuracy for basic geometric primitives and complex relations. **(a)** Evaluation results on LLaVA-8B. **(b)** Evaluation results on Qwen-VL-7B. Across both architectures, primitive-level attributes are more linearly decodable from shallow-to-middle layers, while relation-level attributes reach higher probing accuracy in middle visual layers. The shaded areas highlight the larger probing gains of the merged model on relation-level attributes.

relation-level attributes are relatively weak in shallow layers but increase substantially in middle layers before slightly decreasing in deep layers.

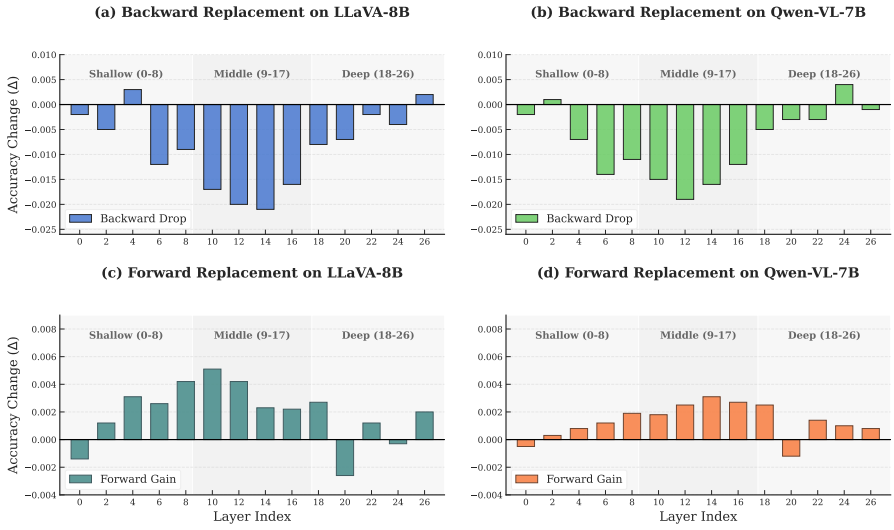


Figure 3: Layer-wise impact of geometry-aware merging on MathVista under two replacement protocols. (a)–(b) **Backward replacement**: replacing merged weights with base weights causes the largest accuracy drops in middle visual layers (9–17). (c)–(d) **Forward replacement**: inserting merged weights into the base model yields the largest gains in the same middle layers. Across both models, under this replacement protocol, the largest changes occur in middle visual layers.

Although relation-level probing accuracy decreases mildly in deep layers, it remains relatively high, suggesting that deep layers continue to preserve relation-sensitive information. We further observe that merging yields only marginal gains on primitive-level attributes but noticeably larger gains on relation-level attributes, especially in middle layers. Overall, these results suggest that geometry-aware merging is particularly beneficial for relation-level geometric understanding. These findings suggest a hierarchical organization of perceptual knowledge in VLMs: early layers capture primitive visual features, while middle layers integrate them into relations. Thus, our merging mainly improves relation representation formation rather than low-level visual feature extraction.

Layer-wise Effects of Replacement. To localize where geometry-aware merging contributes most, we conduct two complementary layer replacement analyses on MathVista (Figure 3). In the first analysis, we start from the merged model and replace one layer at a time with its counterpart from the base model, then measure the resulting accuracy change. In the second analysis, we start from the base model and replace one layer at a time with its counterpart from the merged model, then measure the resulting accuracy change. Together, these two directions provide complementary views of layer-wise contribution from removal-induced degradation and insertion-induced improvement. Across both LLaVA-8B and Qwen-VL-7B, we observe a consistent non-uniform pattern. Replacing shallow or deep layers generally causes only minor performance changes, whereas replacing middle layers consistently produces the largest degradation in the first direction and the largest improvement in the second direction. The consistency of this trend across two backbones and both replacement directions suggests that geometry-aware merging benefits are concentrated in middle layers rather than uniformly distributed across perception layers. Together with the probing results, these findings reveal that perceptual transfer is concentrated in intermediate visual layers. These layers not only provide more accessible representations for relational information, but also have a stronger functional influence on downstream reasoning performance. This suggests that intermediate layers serve as a key interface where transferred perceptual knowledge is converted into task-relevant geometric understanding.

4.4 Further Analysis

The Effect of Scaling Coefficients. We analyze the robustness of perception-side transfer under different merging strengths by varying the coefficient λ from 0.1 to 1.0 with a step size of 0.1, using the PGPS9K-merged Qwen2.5-VL-7B model on MathVista.

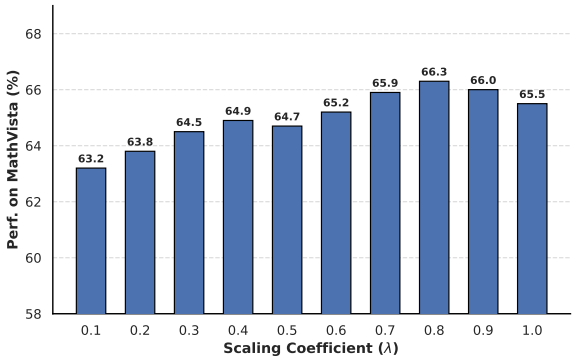


Figure 4: The impact of the scaling coefficient λ

As shown in Fig. 4, performance is relatively low at small λ values, indicating that over-emphasizing τ_{per} while underweighting τ_{vlm} can hurt overall performance. Accuracy increases as λ grows and reaches the highest value (66.3%) at $\lambda = 0.8$, followed by a slight decline at larger values. This trend suggests that balancing the two task vectors is important for stable performance. Since our goal is to analyze perception-side transfer under a fixed merging setup rather than to optimize a benchmark-specific hyperparameter, we use a single coefficient $\lambda = 0.8$ as our default setting.

Comparison across merging methods. We further examine whether the observed effects of perception-side knowledge injection depend on the specific merging algorithm. In addition to linear merging, we evaluate TIES Yadav et al. (2023) and DARE-based variants (DARE-TIES and DARE-Linear) Yu et al. (2024). Following the scaling study, we fix the perception task-vector coefficient to $\alpha_2 = 0.2$. For methods involving parameter sparsification, such as TIES-based merging, we tune the VLM-vector coefficient α_1 within the range $[1, 4]$ to account for the different parameter scales induced by sparsification. As shown in Table 2, different merging strategies lead to different performance trade-offs across benchmarks. Linear merging achieves the highest performance on MathVision, whereas TIES performs better on MathVerse-Vision-Only. The DARE-based variants also yield competitive results in some settings, but do not consistently improve over their non-DARE counterparts. Overall, no single merging strategy dominates across the evaluated benchmarks.

This comparison suggests that the effectiveness of perception-side merging is not unique to linear merging, but can be observed under different parameter-composition strategies. At the same time, the variation across benchmarks indicates that the exact performance gain can depend on how the task vectors are combined. Since our primary goal is to analyze perceptual transfer rather than optimize a merging algorithm, we adopt linear merging as the default intervention. Its simple parameter

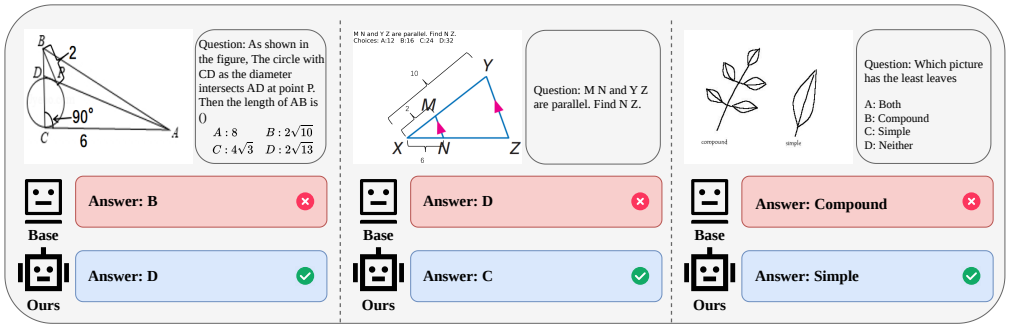


Figure 5: Qualitative comparisons between the base and merged models on three vision-dependent examples. Three cases solved by merging PGPS9K

arithmetic introduces fewer algorithm-specific factors, making the resulting layer-wise changes easier to interpret while still providing competitive downstream performance.

Table 2: Performance comparison across different merging methods on MathVision and MathVerse-Vision-Only.

Method	Weights	MathVision	MathVerse-Vision-Only
Baseline	-	25.4	30.8
Linear	(0.8, 0.2)	26.9	33.1
TIES	(2.8, 0.2)	26.3	35.1
DARE-TIES	(2.0, 0.2)	21.3	34.5
DARE-Linear	(1.8, 0.2)	24.9	32.4

4.5 Case Study

Figure 5 shows three representative examples from vision-dependent subsets of MathVerse. In the first example, the model needs to parse the circle with diameter CD, the intersection point P on AD, and the resulting geometric relations; in the second, it must recognize $MN \parallel YZ$ and use the corresponding proportional structure; the third example is not geometry-specific: the model needs to distinguish the two leaf structures and compare the number of leaves. The base model fails on all three cases, predicting B, D, and Compound, respectively, while the merged model gives the correct answers D, C, and Simple. These examples suggest that merging improves the use of task-relevant visual information in downstream reasoning. While the first two cases involve geometric perception, the third case further suggests that the benefit may extend beyond geometry-specific cues to more general visual discrimination. More cases are provided in Appendix C.3.

5 Conclusion

In this paper, we study geometry-aware perceptual transfer in VLMs through perception-side model merging. By merging geometry-aware task vectors into the vision encoder and projector, we focus on perception-side parameter changes while keeping the LLM fixed. Experiments across multiple VLMs and visual mathematical benchmarks show significant improvements, especially on geometry-sensitive and vision-dependent tasks. Further analyses suggest that relation-aware geometric supervision plays an important role in these gains. Layer-wise studies further reveal that primitive cues emerge in shallow layers, while relation-sensitive information is more prominent in middle layers. Overall, our study provides a layer-wise view of how perceptual knowledge is represented and transferred within VLMs. This perspective may help guide more targeted perception-side adaptation by identifying the representations and layers most relevant to structured visual understanding. An important future direction is to examine whether similar layer-wise organization also emerges for perceptual knowledge beyond geometry.

AI USE STATEMENT

In this work, we used generative AI tools to assist with coding and manuscript writing. Generative AI was not used to produce conclusions unsupported by objective experimental evidence. All AI-assisted content, including but not limited to code and manuscript text, was manually reviewed and verified by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed model configurations, task vector construction procedures, training settings, merging hyperparameters, evaluation protocols, and layer-wise analysis settings in the main paper and appendix. We additionally specify the generation procedure, random seed, data splits, and training configuration for the synthetic probing dataset. Potential overlaps between task-vector training data and downstream evaluation benchmarks are detected and removed to prevent data leakage. The code for perception-side tuning, model merging, probing, and evaluation, together with the synthetic data generation scripts, will be publicly released upon acceptance.

REFERENCES

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.06786*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Shihao Cai, Keqin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, and Bo Zheng. Geogpt4v: Towards geometric multi-modal large language models with geometric image generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 750–766, 2024.

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.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3313–3323, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.218. URL https://aclanthology.org/2022.emnlp-main.218/.

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.

Mingrui Chen, Haogeng Liu, Hao Liang, Huaibo Huang, Wentao Zhang, and Ran He. Unlocking the potential of difficulty prior in rl-based multimodal reasoning. *arXiv preprint arXiv:2505.13261*, 2025a.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. Bring reason to vision: Understanding perception and reasoning through model merging. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=ntCAP6tMoX.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. MINT-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=vMpvtSmtXY.

Junhao Dong, Xinghua Qu, Cong Zhang, Sua Qi Rong, Nguyen Duc Thai, Wenbo Pan, Xinfeng Li, Tongliang Liu, Piotr Koniusz, and Yew-Soon Ong. Tug-of-war no more: Harmonizing accuracy and robustness in vision-language models via stability-aware task vector merging. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KOO1cDm2bt.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. In *International Conference on Learning Representations*, volume 2025, pp. 3490–3511, 2025.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18695–18705, 2025. doi: 10.1109/CVPR52734.2025.01742.

Yihan Hao, Mingliang Zhang, Fei Yin, and Lin-Lin Huang. Pgdp5k: A diagram parsing dataset for plane geometry problems. In *2022 26th international conference on pattern recognition (ICPR)*, pp. 1763–1769. IEEE, 2022.

Minbin Huang, Runhui Huang, Chuanyang Zheng, Jingyao Li, Guoxuan Chen, Han Shi, and Hong Cheng. Answer-consistent chain-of-thought reinforcement learning for multi-modal large langauge models. *arXiv preprint arXiv:2510.10104*, 2025.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Zheng Jiang, Yiming Chen, Nan He, Jiahui Chen, Chaoyang Li, Houde Qian, and Lifeng Sun. Test-time scaling over perception: Resolving the grounding paradox in thinking with images. In *Proceedings of the ACM Conference*, 2026.

Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018.

Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning, 2018. URL https://openreview.net/forum?id=SyunbfbAb.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yufang Liu, Yao Du, Tao Ji, Jianing Wang, Yang Liu, Yuanbin Wu, Aimin Zhou, Mengdi Zhang, and Xunliang Cai. The role of visual modality in multimodal mathematical reasoning: Challenges and insights. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22596–22611, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1102. URL https://aclanthology.org/2025.acl-long.1102/.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 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.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jian-Guang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mMPMHWOdOy.

Ruilin Luo, Zhuofan Zheng, Lei Wang, Yifan Wang, Xinzhe Ni, Zicheng Lin, Songtao Jiang, Yiyao Yu, Chufan Shi, Ruihang Chu, Jin zeng, and Yujiu Yang. Unlocking multimodal mathematical reasoning via process reward model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=96I8PGPALv.

Jingkun Ma, Runzhe Zhan, Yang Li, Di Sun, Hou Pong Chan, Lidia S. Chao, and Derek F. Wong. Visaidmath: Benchmarking visual-aided mathematical reasoning. In *1st Workshop on VLM4RWD @ NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=frp6TqqcTF.

Ziyu Ma, Chenhui Gou, Yiming Hu, Yong Wang, Bohan Zhuang, and Jianfei Cai. Where and what matters: Sensitivity-aware task vectors for many-shot multimodal in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7892–7900, 2026.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

Brandon Ong, Tej Deep Pala, Vernon Toh, William Chandra Tjhi, and Soujanya Poria. Training vision-language process reward models for test-time scaling in multimodal reasoning: Key insights and lessons learned. *arXiv preprint arXiv:2509.23250*, 2025.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, Yifan Zhang, 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 (Volume 1: Long Papers)*, pp. 20023–20070, 2025.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4663–4680, 2024.

Trieu Trinh, Yuhuai Wu, Quoc Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 2024. doi: 10.1038/s41586-023-06747-5.

Jinping Wang, Zhiqiang Gao, Zhang Dinggen, and Zhiwu Xie. Escaping optimization stagnation: Taking steps beyond task arithmetic via difference vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26310–26318, 2026.

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. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=QWTCcxMpPA.

Wenjun Wu, Lingling Zhang, Jun Liu, Xi Tang, Yaxian Wang, Shaowei Wang, and Qianying Wang. E-gps: Explainable geometry problem solving via top-down solver and bottom-up generator. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13828–13837, 2024. doi: 10.1109/CVPR52733.2024.01312.

Peijin Xie, Zhen Xu, Bingquan Liu, and Baoxun Wang. M^3-ace: Rectifying visual perception in multimodal math reasoning via multi-agentic context engineering. *arXiv preprint arXiv:2603.08369*, 2026.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xtaX3WyCj1.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8):1–41, 2026a.

Tianyu Yang, Sihong Wu, Yilun Zhao, Zhenwen Liang, Lisen Dai, Chen Zhao, Minhao Cheng, Arman Cohan, and Xiangliang Zhang. A survey of multimodal mathematical reasoning: From perception, alignment to reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42386–42408, 2026b.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*. PMLR, 2024.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. In *International Conference on Learning Representations*, volume 2024, pp. 40320–40341, 2024.

Jiaxin Zhang and Yashar Moshfeghi. GOLD: Geometry problem solver with natural language description. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 263–278, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.19. URL https://aclanthology.org/2024.findings-naacl.19/.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1859–1869, 2025a. URL https://api.semanticscholar.org/CorpusID:277066741.

Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. A multi-modal neural geometric solver with textual clauses parsed from diagram. In *IJCAI*, 2023.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Shanghang Zhang, Gao Peng, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. In *International Conference on Learning Representations*, volume 2025, pp. 87955–87989, 2025b.

Shan Zhang, Aotian Chen, Yanpeng Sun, Jindong Gu, Yi-Yu Zheng, Piotr Koniusz, Kai Zou, Anton van den Hengel, and Yuan Xue. Primitive vision: Improving diagram understanding in mllms. 2025c.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 48337–48383, 2025.

A Experiment Details

A.1 Datasets

Training Datasets. In this paper, we utilize four geometry-centric datasets to obtain task vectors.

- **PGPS9K** Zhang et al. (2023): The Plane Geometry Problem Solving dataset, developed by CASIA, contains 9,022 geometry problems paired with 4,000 non-duplicate geometry diagrams. It features fine-grained diagram annotations translated into structural and semantic clauses, alongside interpretable solution programs. For our experimental setup, we adopt its standard partition consisting of 8,433 training samples and 589 distinct testing samples.
- **Geo3K** Lu et al. (2021): Also referred to as Geometry3K, this textbook-derived dataset consists of 3,002 multiple-choice geometry problems. It serves as a foundational benchmark for testing dense diagram-text alignment and multi-step geometric calculation within secondary school mathematics curricula.
- **UniGeo** Chen et al. (2022): A large-scale unified geometric dataset containing over 13,000 problems. It formally integrates both algebraic calculation tasks and rigorous logic-based geometric proof tasks, providing a comprehensive evaluation of a model's dual capabilities in numerical and symbolic geometric reasoning.
- **PGDP5K** Hao et al. (2022): The Plane Geometry Diagram Parsing dataset contains 5,000 high-resolution geometric image samples. It is meticulously annotated with dense, primitive-level geometric entities (such as points, lines, and circles) and their corresponding relational topology graphs, focusing explicitly on parsing visual geometric structures.

We strictly adhere to the official data usage and open-source terms of all utilized assets. Specifically, **PGPS9K** and **PGDP5K** are released by the Institute of Automation of the Chinese Academy of Sciences (CASIA) under a Custom Academic Agreement, which grants free access exclusively for non-commercial academic research purposes. **Geometry3K** is publicly distributed under the permissive **MIT License** within its official repository. For **UniGeo**, the code and data are made publicly available by the authors for research purposes, and all experiments in this work strictly restrict its usage to non-commercial scientific exploration.

Evaluation Datasets. In this paper, we evaluate the merged model on these benchmarks:

- **MathVista** Lu et al. (2024): A comprehensive and large-scale benchmark designed to assess mathematical reasoning in diverse visual contexts. It integrates a wide array of mathematical tasks, challenging the model's ability to tackle complex, multi-step reasoning problems accompanied by rich visual information.
- **MathVerse** Zhang et al. (2024): An evaluation benchmark rigorously designed to probe whether VLMs genuinely perceive visual diagrams or merely guess based on textual clues.
- **MMStar** Chen et al. (2024): An elite vision-language benchmark constructed by strictly filtering out visually-irrelevant samples. It specifically evaluates true multimodal capabilities, ensuring that the models must genuinely rely on geometric and visual elements, rather than language priors, to solve the mathematical problems.
- **DynaMath** Zou et al. (2025): A dynamic visual mathematical benchmark tailored to assess the robustness of multimodal reasoning. By applying programmatic visual and textual transformations to root problems, it systematically evaluates whether models can maintain their geometric perception and reasoning capabilities under varying multimodal perturbations.
- **MathVision** Wang et al. (2024): A highly challenging benchmark consisting of visual mathematical problems directly sourced from real-world mathematics competitions. It is utilized to test the upper limits of a model's advanced geometric intuition and complex analytical skills in authentic, olympiad-level visual contexts.

We strictly adhere to the specific licensing terms and usage conditions of each evaluation benchmark. For **MathVista**, its novel contributions (e.g., new subsets and metadata) are distributed under the CC

BY-SA 4.0 license, and we strictly utilize this dataset exclusively as a test set to comply with its explicit prohibition against model training. The source code and evaluation framework of **MathVerse** are hosted under the permissive MIT License, and we restrict the use of its aggregated multimodal data strictly to academic research. Regarding **MMStar**, the benchmark is made publicly available by the authors for research purposes, which we utilize exclusively for non-commercial scientific benchmarking while fully respecting the copyright of the source images. Similarly, the evaluation code and dynamic test sets of **DynaMath** are publicly accessible, and all experiments in this work restrict its usage to non-commercial academic evaluation. Finally, **MathVision** is sourced directly from real-world mathematics competitions and is publicly available for research, and we strictly limit its application to academic testing purposes.

B Probing Dataset and Experimental Details

To analyze the layer-wise accessibility of geometric representations within the vision encoders of large vision-language models (VLMs), we constructed a lightweight, controlled synthetic geometry probing dataset. This dataset is designed to minimize the influence of textual priors from the large language model (LLM) and to probe the accessibility of geometric primitives and relations at different network depths.

Table 3: Probing Dataset Statistics and Metrics

Task	**Label Space**	**Samples**	**Class Balance**	**Metric**
Line presence	{yes, no}	1,000	500 / 500	Accuracy
Angle type	{acute, right, obtuse}	1,500	500 / 500 / 500	Accuracy
Line relation	{parallel, perpendicular, intersect}	1,500	500 / 500 / 500	Accuracy
Intersection count	{0, 1, 2plus}	1,500	500 / 500 / 500	Accuracy
Total	-	5,500	Balanced per class	-

Table 4: Unified Prompt Templates

Task	**Prompt**	**Allowed Output Set**
Line presence	Is there a straight line in the image? Answer yes or no.	{yes, no}
Angle type	What type of angle is shown? Answer one of: acute, right, obtuse.	{acute, right, obtuse}
Line relation	What is the relation between the two main lines? Answer one of: parallel, perpendicular, intersect.	{parallel, perpendicular, intersect}
Intersection count	How many intersections are there? Answer one of: 0, 1, 2plus.	{0, 1, 2plus}

B.1 Data Generation and Protocols

We construct a synthetic dataset comprising four attribute-prediction tasks: line presence, angle type, line relation, and intersection count. All images are rendered at 256×256 resolution with a white background and black geometric strokes to eliminate background noise.

Data generation follows deterministic rules with a fixed random seed (`seed=42`):

- **Line presence:** Generates either one straight line (yes) or 2–4 hollow distractor circles (no). The distractors prevent pixel-density shortcut learning.
- **Angle type:** Two rays from a common center, with ranges acute $\in [20^\circ, 70^\circ]$, right $\in [88^\circ, 92^\circ]$, and obtuse $\in [110^\circ, 160^\circ]$.
- **Line relation:** Two main lines generated as parallel, perpendicular, or intersecting, with small random geometric perturbations (e.g., $\pm 3^\circ$).
- **Intersection count:** Mapped to 0 (parallel), 1 (single intersection), or 2plus (complex three-line pattern).

Each class contains 500 samples, yielding 5,500 samples in total, serialized into a single JSONL file.

B.2 Dataset Statistics and Prompt Templates

For probing, we perform a stratified split per task with ratios 0.7/0.1/0.2 (train/val/test). We use closed-set answer spaces and exact-match (EM) evaluation.

Note on Input Modality: To comply with the standard multimodal interface of VLMs, the prompt text is provided alongside the images. However, the linear probes extract features **exclusively** from the vision encoder layers, ensuring geometric attributes are decoded purely from visual representations.

B.3 Linear Probe Training and Evaluation Settings

During training, the VLM's vision encoder weights are kept completely frozen. For each visual layer, we train a separate linear probe (single-layer classifier with bias).

- **Feature Extraction:** Given a layer output tensor, we reshape it to $[B, -, D]$ and perform mean-pooling over non-batch and non-hidden dimensions to obtain a $[B, D]$ global feature representation.
- **Hyperparameters:** We utilize the AdamW optimizer with a learning rate of `1e-2` and a weight decay of `1e-4`. This learning rate optimally accelerates convergence within our standardized optimization space.
- **Regularization & Stopping:** Models are trained for up to 200 epochs. An early-stopping mechanism (patience = 20) based on validation accuracy is employed to prevent overfitting.
- **Normalization:** To strictly prevent data leakage, layer-wise Z-score normalization is computed *solely* from the training split statistics and subsequently applied to the train, validation, and test splits. The test accuracy is reported as the primary metric.

Table 5: Ablation on Feature Pooling Strategies (Best-Layer Test Accuracy)

Task	Mean Pooling (Ours)	CLS-token	Max Pooling
Line presence	94.00%	92.50%	93.50%
Angle type	93.33%	92.67%	93.00%
Line relation	75.33%	76.00%	74.67%
Intersection count	77.33%	77.33%	78.00%
Average	**84.18%**	**83.91%**	**84.00%**

Robustness to Learning Rate Selection. In our main experiments, we employ an AdamW optimizer with a learning rate of 1×10^{-2}. While larger than typical default values, this choice accelerates convergence without sacrificing performance. To empirically justify this, we performed a learning rate sweep over $\{1 \times 10^{-3}, 3 \times 10^{-3}, 1 \times 10^{-2}\}$ on the full probing dataset using the Qwen vision encoder. As shown in Table 6, the 1×10^{-2} setting achieves a slightly superior average best-layer test accuracy (84.18%).

To verify optimization stability, we further plot the validation-loss convergence curves for all four probing tasks (Figure 6), where each curve averages the validation loss over all visual layers. The results demonstrate that AdamW with an $lr = 1 \times 10^{-2}$ does not exhibit divergence or persistent oscillation. Although its initial loss is sometimes higher, it consistently converges stably to lower or comparable final validation losses alongside smaller learning rates.

Table 6: Learning Rate Sweep for Linear Probes (Qwen Perception Module)

Learning Rate	Avg. Best-Layer Test Accuracy
1×10^{-3}	83.82%
3×10^{-3}	84.09%
1×10^{-2} **(Ours)**	**84.18%**

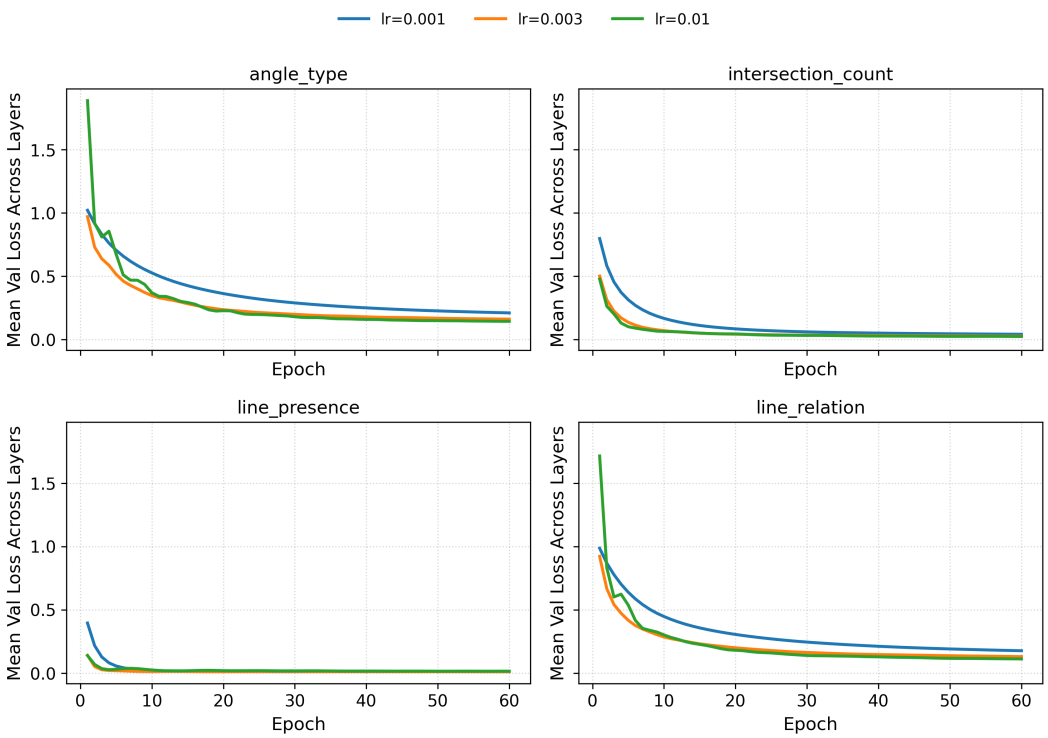


Figure 6: Validation loss convergence curves across four probing tasks. Each curve represents the validation loss averaged over all visual layers. The $lr = 1 \times 10^{-2}$ setting maintains stable optimization without divergence.

Absolute Independence from Textual Prompts. To eliminate concerns that textual prompts might leak language priors into the probed representations, we explicitly verified the structural independence of the perception module in our setup. We conducted a direct tensor-level comparison of the visual features extracted using task-specific prompts versus a neutral visual prompt. The maximum absolute difference across all visual layers is mathematically strict: $\text{max_abs_diff} = 0.0$. Consequently, the entire layer-wise probe curves are mathematically identical regardless of the text condition. This empirically confirms that our linear probes are trained on *pre-fusion* visual representations, which are physically isolated from and completely uninfluenced by the textual input modality.

Insensitivity to Feature Pooling Strategies. A typical concern regarding mean-pooling is that it might artificially discard highly localized spatial cues (e.g., specific geometric intersection points). To address this, we compared mean pooling against CLS-token extraction and max pooling for aggregating per-layer token features. As detailed in Table 5, the empirical results demonstrate that all three aggregation strategies yield highly consistent best-layer probe accuracies and exhibit the identical hierarchical trend (i.e., primitives established in shallow layers, while complex relations peak in middle layers). Because our controlled synthetic dataset features clean, minimal strokes on a pure white background, topological primitives are globally readable across the visual patch sequence. Therefore, our core mechanistic conclusions are robust and not skewed by the specific token aggregation rule.

B.4 Geometry-aware Supervision vs. Weakened Supervision

Our main experiments construct perception task vectors using geometry-aware visual descriptions, following the diagram-to-description supervision style. These descriptions include visible geometric primitives, symbolic annotations, and geometric relations, and are used only as training targets rather than inference-time inputs. To examine whether the improvement comes from explicit relation-

aware supervision instead of generic geometry-domain textual supervision, we construct a relation-weakened supervision baseline.

This experiment follows the same setting as the main experiments, including the training images, training split, optimization configuration, merging coefficient, and evaluation protocol. The only difference lies in the supervision target used during perception tuning. In the full geometry-aware setting, the target includes structural topology together with explicit semantic relational annotations, such as geometric relation predicates and their grounding/alignment to visual symbols and geometric entities. In the relation-weakened setting, we retain perceptual-level visual evidence, including observable geometric primitives, point labels, line structures, algebraic expressions, and visual symbols, while removing explicit high-level relational supervision. Specifically, annotations for parallelism, equality, perpendicularity, and angle-measure correspondences are not provided as structured semantic labels.

We note that semantic cues may still implicitly exist in the diagram itself, such as through visual markers or algebraic symbols. Therefore, this baseline does not remove all semantic information from the image. Instead, it removes explicit relational supervision from the textual target, allowing us to test whether structured relation annotations are important for constructing transferable perception task vectors.

Table 7: Effect of geometry-aware supervision compared with weakened supervision. Full denotes Full geometry-aware supervision; Weakened denotes Weakened supervision

Method	MathVista	V-I	V-D	V-O
Baseline	65.5	34.9	36.0	30.8
Weakened	65.4	35.4	37.1	32.9
Full	66.3	37.5	39.4	33.1

As shown in Table 7, relation-weakened supervision improves several vision-dependent subsets over the baseline, suggesting that perceptual-level annotations such as primitives, symbols, and line structures already provide useful visual signals. However, full geometry-aware supervision consistently outperforms the relation-weakened variant across all evaluated metrics. In particular, it improves over relation-weakened supervision by 2.1 points on V-I and 2.3 points on V-D. These results suggest that the benefit of the perception task vector is not merely due to exposure to geometry-domain textual descriptions. Instead, explicit relation predicates and their grounding to geometric entities provide additional transferable supervision for geometry-sensitive visual mathematical reasoning.

Full geometry supervision

<Visual_Evidence>

Based on the image, the geometric topology and semantics are as follows:
Structural Topology

- line BA
- line BCE
- line ADE
- line CD

Semantic Annotations

- $BA \parallel CD$
- $m\angle DAB = 68$
- $m\angle EDC = 2x$
- $m\angle ABC = 3x - 15$
- $m\angle CED = y^2$

Note: This is a pure visual observation. No mathematical reasoning is performed.

</Visual_Evidence>

Weakened geometry supervision

<Visual_Evidence>

Based on the image, the visible geometric primitives and symbols are as follows:
Structural Topology

- line BA
- line BCE
- line ADE
- line CD

Visual Symbols

- numeric token: 68
- algebraic token: $2x$
- algebraic token: $3x - 15$
- algebraic token: y^2
- point labels: A, B, C, D, E

Note: This is a pure visual observation. No mathematical reasoning is performed.

</Visual_Evidence>

C More Experiments

C.1 Merging vs. Pure Fine-tuning.

We further compare linear merging with a pure fine-tuning baseline. Here, pure fine-tuning means directly using the fine-tuned model for inference, without any post-hoc parameter merging with the original VLM. As shown in Table 8, merging consistently outperforms pure fine-tuning on both backbones. For Qwen-VL-7B, merging achieves 66.3%, substantially higher than 60.7% from pure fine-tuning. For LLaVA-8B, merging also yields a small but consistent gain (38.5% vs. 38.1%). These results suggest that linear merging is a more effective default strategy than directly using the fine-tuned model.

Table 8: Performance comparison between pure fine-tuning and linear merging on MathVista.

Model	Pure fine-tuning (%)	Merging (%)
Qwen-VL-7B	60.7	**66.3**
LLaVA-8B	38.1	**38.5**

C.2 Analysis on Layer-wise Sensitivity

We conduct an additional study to evaluate the sensitivity across different layers. Specifically, we adopt a "uniform $1/N$ masking" technique. In this setup, we replace the pretrained weights of a targeted visual layer with a uniform constant matrix, where each parameter is rigidly assigned a value of $1/N$ (with N defined as the size of the first dimension of the weight matrix).

Figure 7 reveals a highly consistent trend across both LLaVA-8B and Qwen-VL-7B backbones. We observe that perturbing the parameter spaces in the shallow layers leads to the most catastrophic performance degradation. This destructive influence remains highly considerable throughout the middle layers. In contrast, while the performance continues to drop when masking deeper layers, the magnitude of degradation becomes noticeably smaller, indicating a higher degree of resilience at the end of the network.

This reverse-verification experiment firmly reinforces our primary findings: the mathematical perception capabilities of VLMs are exceptionally sensitive to perturbations in the early-to-middle parameter spaces. It perfectly aligns with the conclusion that the shallow and middle layers act as the

critical functional bottleneck for geometric problem-solving, whereas the deeper layers are relatively more robust to structural disruption.

C.3 More Cases

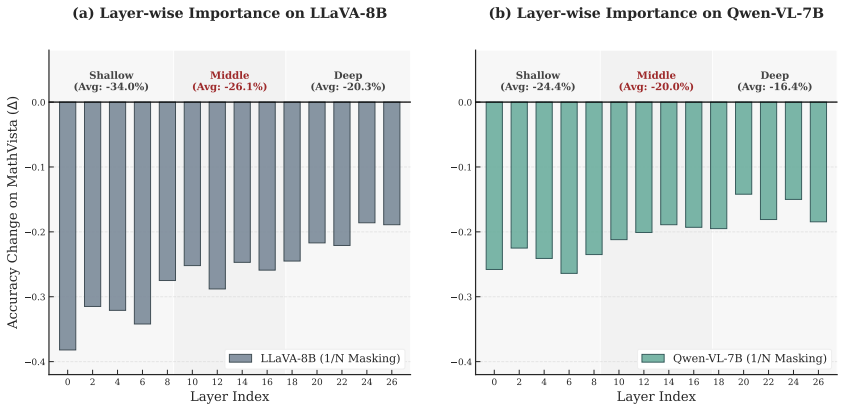


Figure 7: Measuring layer-wise parameter importance via uniform $1/N$ weight masking. **(a)** Results on LLaVA-8B. **(b)** Results on Qwen-VL-7B. To completely avoid overlap with long vertical bars, interval text descriptions are positioned above the 0.0 baseline. Perturbing parameter spaces in shallow and middle layers triggers severe performance degradation, while deep layers exhibit higher resilience.

To provide a more intuitive understanding of the perceptual enhancements brought by our geometry-aware model merging, we exhibit four representative case studies harvested from the evaluation benchmarks. These qualitative comparisons highlight the distinct gap in geometric reasoning between the baseline and our merged models.

In the first case of circular geometry reasoning (Figure 8), given a circle where $AB = AC$ and $\angle BAC = 70°$, the task requires calculating the inscribed angle $\angle AEC$. The baseline model settles on an incorrect guess of $75°$, whereas the merged model successfully perceives the geometric property that inscribed angles subtending the same arc are equal ($\angle AEC = \angle ABC$), correctly deriving the true answer as $55°$ via isosceles triangle properties.

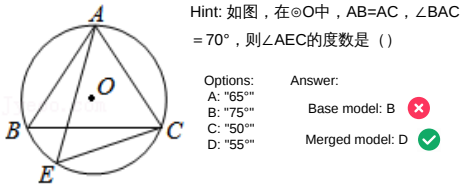
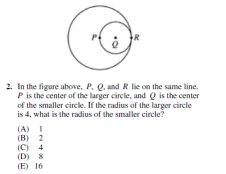


Figure 8: Case study comparing base model and merged model on circular geometry reasoning.

In the second parallelogram scenario (Figure 9), the model must solve for $\angle COD$ given two local angle annotations ($42°$ and $23°$). Lacking fine-grained spatial coordination, the baseline model fails to locate the correct target and outputs an incorrect option, while the merged model successfully leverages alternate interior angles and triangle angle-sum theorems to output the exact solution of $65°$.

For the third task involving concentric and intersecting circles (Figure 10), P is the center of the larger circle and Q is the center of the smaller one. When asked to deduce the radius of the smaller circle given the larger radius is 4, the baseline model exhibits severe multimodal hallucination and outputs an irrelevant degree value ($40°$), whereas the merged model accurately grounds the spatial alignment on the line segment, recognizing that the radius of the larger circle forms the diameter of the smaller circle to output the correct mathematical solution.

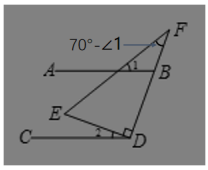


Hint: In the figure above, P, Q, and R lie on the same line. P is the center of the larger circle, and Q is the center of the smaller circle. If the radius of the larger circle is 4, what is the radius of the smaller circle?

Options:
A: "20°"
B: "25°"
C: "30°"
D: "40°"

Answer:
Base model: D
Merged model: A

Figure 9: Case study comparing base model and merged model on structural relationship of intersecting circles.



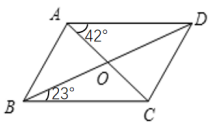
Hint: As shown in the figure, the straight line AB parallel CD, then the degree of angle 2 is ()

Options:
A: "20°"
B: "25°"
C: "30°"
D: "40°"

Answer:
Base model: D
Merged model: A

Figure 10: Case study comparing base model and merged model on parallel line geometric constraints

Finally, in the fourth angular problem involving parallel lines (Figure 11), where $AB \parallel CD$, the model is required to compute $\angle 2$ using the programmatic angle label $70° - \angle 1$. The baseline model provides a wrong prediction due to text-diagram misalignment, while our merged model reliably decodes the parallel spatial orientation and geometric constraints to select the true option. In summary, these qualitative comparisons firmly demonstrate that our perception-merging paradigm enables the VLM to transition from text-biased hallucinations to genuine, structure-grounded visual geometric perception.



Hint: In parallelogram ABCD, then angle COD is ()

Options:
A: "61°"
B: "63°"
C: "65°"
D: "67°"

Answer:
Base model: B
Merged model: C

Figure 11: Case study comparing base model and merged model on parallelogram angular derivation.