PASA: POST-MERGE PERCEPTION–REASONING ASYMMETRY AS A SELF-ALIGNMENT SIGNAL FOR MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging offers a direct way to transfer specialized reasoning capabilities to multimodal large language models (MLLMs), but successful transfer must also preserve their ability to ground answers in visual evidence. We find that merging a reasoning-specialized language model into an MLLM improves reasoning performance while degrading fine-grained visual perception, a phenomenon we term post-merge perception–reasoning asymmetry. We investigate whether the reasoning capability retained after merging can itself provide supervision for recovering visual perception. Comparisons on the same self-alignment queries show that reasoning-prompted responses are more accurate than perception-prompted responses, suggesting a source of supervision within the model. Building on this observation, we introduce PASA, a two-stage framework for post-merge self-alignment. PASA first learns to generate structured visual-evidence traces through supervised fine-tuning on synthetic data. It then applies asymmetric GRPO: a reasoning-prompted role supplies detached answer anchors, while perception-prompted trajectories are optimized for answer agreement and structural validity. Both roles share parameters, allowing anchors to be regenerated from the updated model without ground-truth answer labels during reinforcement learning. Across model scales and merging methods, PASA recovers much of the post-merge perception loss while further improving reasoning. Evaluations of generated visual evidence show improved grounding, and stage-wise self-training comparisons support the benefit of asymmetric supervision. These results demonstrate how capabilities retained after merging can serve as an internal learning signal for post-merge self-alignment.

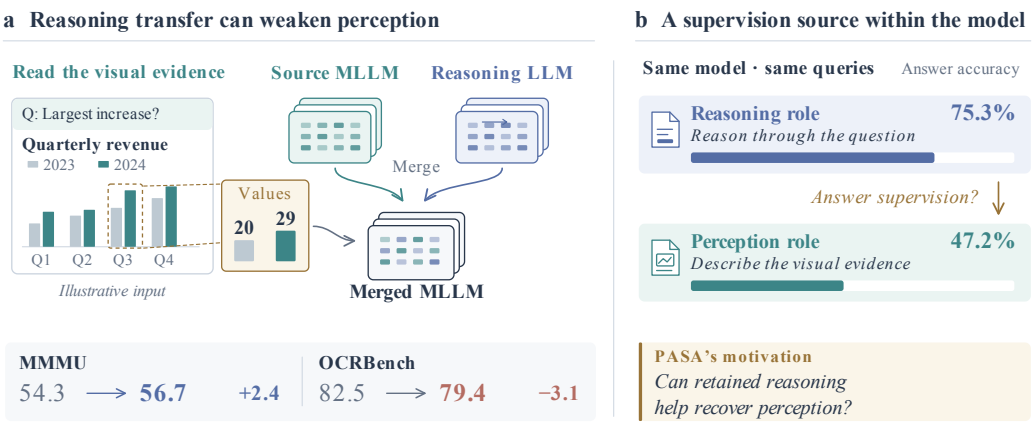


Figure 1: **Post-merge asymmetry motivates an internal source of supervision. a**, Illustrative visual evidence and measured source-to-merge scores in the primary 2B setting. The chart is not a model output. **b**, Same-query answer accuracy of two prompted roles at the stabilized checkpoint before RL (6,000 training queries). The gap motivates reasoning-answer supervision; an answer anchor does not directly verify visual evidence. Values summarize reported aggregates, not interval estimates.

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.

1 INTRODUCTION

Multimodal reasoning requires a model to extract visual evidence and draw conclusions from it. Answering a question about a chart, for example, depends on identifying the relevant values and their relationships as well as performing the subsequent calculation. Model merging offers a way to compose specialized checkpoints without jointly retraining their source models (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023). Its application to language and multimodal models provides a route to transferring specialized reasoning capabilities into an existing vision-language system (Zhou et al., 2024; Wei et al., 2025c; Wei & Chen, 2025b). For this transfer to improve multimodal reasoning, however, the resulting model must retain the visual capabilities on which its answers depend.

We study this requirement when merging reasoning-specialized language models with compatible MLLMs. Across the model pairs and merging methods evaluated, we observe a recurring imbalance: reasoning-oriented benchmark scores improve, while performance on fine-grained visual tasks declines. The affected capabilities include reading image text, interpreting charts, and resisting visually unsupported answers. We refer to this pattern as *post-merge perception–reasoning asymmetry*. Figure 1a illustrates how a merged model can gain reasoning capability while losing part of the visual grounding inherited from its source MLLM. Successful capability composition therefore requires examining not only what the model gains, but also what it retains.

This observation motivates a question: can the reasoning capability retained after merging provide supervision for recovering visual perception? We examine this possibility through two prompted roles of the same model. The reasoning role reasons through the question to produce an answer, whereas the perception role explicitly describes visual evidence before answering. On the same multimodal queries, the reasoning role produces more accurate answers at the checkpoint preceding reinforcement learning (Figure 1b). This comparison identifies a potential supervisory source within the model: answers from its more reliable reasoning behavior can guide the learning of its perception-oriented trajectories. The opportunity lies in how the shared model is prompted and trained, without requiring a separate model to supply the answer targets.

We introduce PASA, a two-stage framework for post-merge self-alignment. The first stage, *Perception Action-Space Stabilization*, uses supervised fine-tuning on synthetic traces to establish structured visual-evidence generation. This gives the model a consistent form in which to express and sample visual observations. The second stage, *Detached-Anchor Asymmetric GRPO*, uses answers from the reasoning role as anchors for optimizing perception-role trajectories. Rewards combine agreement with the anchor answer and structural validity. Each anchor is held fixed within an update, while policy gradients are computed from the perception trajectories. Because both roles share parameters, subsequent anchors are generated from the updated model. PASA thus directs supervision between different behaviors of one evolving model, without ground-truth answer rewards during reinforcement learning.

Across 2B, 8B, and 14B models and multiple merging methods, PASA recovers much of the post-merge perception loss while further improving reasoning. In the primary 2B setting, OCRBench increases from 79.4 to 83.2 and MathVista from 63.4 to 69.5 relative to direct merging. Evaluations of generated visual evidence also show improved grounding. With the stabilization stage held fixed, comparisons on the same queries favor reasoning-role anchors over symmetric self-training, supporting the evaluated cross-role construction. Removing direct supervision on final-answer tokens during stabilization preserves nearly all of the completed pipeline's gains. Together, these results support using capabilities retained after merging as an internal learning signal for post-merge self-alignment.

Our contributions are threefold:

- We characterize post-merge perception–reasoning asymmetry across evaluated model scales and merging methods, and identify a same-query difference in role reliability that can support self-alignment.
- We introduce PASA, which combines structured perception-trace learning with detached reasoning-role anchors to optimize perception trajectories within a shared model.
- We demonstrate perception recovery and reasoning gains, supported by visual-evidence evaluation and stage-wise comparisons of the supervision used during post-merge training.

2 Related Work

Model merging for capability composition. Model merging composes capabilities by averaging weights, merging task vectors, or resolving parameter interference (Wortsman et al., 2022; Ilharco et al., 2023; Matena & Raffel, 2022; Yadav et al., 2023; Yu et al., 2024; Jin et al., 2023; Wei et al., 2025d). Recent work extends this idea to LLMs and multimodal models, showing that parameter-level composition can transfer skills without joint training (Zhou et al., 2024; Wei et al., 2025c; Cheng et al., 2025; Wei & Chen, 2025b; Chen et al., 2025). These methods primarily ask how to preserve or add capabilities. Our focus is complementary: we study a cross-modal failure mode in which merging a reasoning-specialized LLM into an MLLM improves reasoning-oriented scores while degrading fine-grained perception in our evaluated settings, and we exploit this asymmetry as a training signal.

Multimodal reasoning and reinforcement learning. Long-chain reasoning and RL have become central to recent LLM and MLLM development (Wei et al., 2022; OpenAI, 2024; Yao et al., 2024; Shao et al., 2024; Guo et al., 2025; Schulman et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023; Yu et al., 2025b). In multimodal settings, recent approaches improve mathematical or visual reasoning through supervised traces, rule-based rewards, or cold-start RL (Xu et al., 2024; Liu et al., 2025; Huang et al., 2025; Zhang et al., 2025; Wei et al., 2025b; Xia et al., 2025; Li et al., 2025; Yu et al., 2025a; Thawakar et al., 2025). These approaches typically rely on verified answer labels, curated reasoning traces, or auxiliary perceptual tasks. PASA targets a different post-merge regime: during Stage-II RL, rewards are derived from the same model's detached reasoning role rather than from ground-truth answer labels. Self-training uses model-generated answers as supervision. MM-UPT (Wei et al., 2025a) forms majority targets from same-policy samples, whereas PC-GRPO (Jeddi et al., 2025) and Visual Jigsaw (Wu et al., 2025) use programmatically verifiable visual proxy tasks. PASA instead uses a reliability gap observed after stabilization: reasoning-role answers supervise perception-role trajectories of a shared merged policy. Its distinction is cross-role supervision on target queries rather than same-role consensus or auxiliary visual tasks, not a new policy optimizer. The measured gap does not establish that merging alone caused it. Image and video captioning express visual content in language (Vinyals et al., 2015; Wei & Chen, 2025a; Hu et al., 2019); OCRBench and HallusionBench diagnose visual-text understanding and hallucination (Liu et al., 2024; Guan et al., 2024). Multimodal reasoning research studies how perception supports inference (Liang et al., 2025). PASA learns a structured output space for cross-role optimization, not a hard decoding constraint or visual-fact verifier. Trace controls and a separate grounding audit evaluate this design.

3 PASA: Cross-Role Self-Alignment

PASA uses the reasoning behavior retained by a merged model as a source of answer-level supervision for learning explicit visual-evidence trajectories.

3.1 Cross-Role Learning in a Shared Model

Let $Q = (I, q)$ be an image–question pair and π_θ the merged MLLM. The prompts $\mathcal{P}_{\text{reason}}$ and $\mathcal{P}_{\text{perc}}$ elicit reasoning followed by an answer, and an explicit visual-evidence caption followed by an answer, respectively. These are different conditional behaviors of one policy, not separately parameterized networks. We write a perception trajectory as $y_p = (e, a)$, where e is the structured caption and a is its final answer.

The supervision direction is motivated by a same-query comparison at the stabilized primary 2B checkpoint, after Stage I and before Stage-II RL. On the same 6,000 Stage-II training queries, reasoning-role and perception-role answer accuracies are 75.3% and 47.2%, respectively (Figure 1b). Gold answers are used only for this offline diagnosis, not for selecting anchors or computing RL rewards. This within-checkpoint gap motivates using reasoning answers to guide perception-oriented learning; it does not establish that merging alone caused the gap.

The Stage-II supervisory link uses only the *extracted answer*: the perception trajectory is not trained to reproduce the reasoning role's full response. This separates the source of the supervisory target from the behavior being optimized. PASA implements this relation in two stages (Figure 2). Stage I teaches structured visual-evidence generation (a–b), establishing the output form to be optimized.

Stage II compares sampled perception trajectories against reasoning-role answers (c–d), providing a cross-role training signal within that form.

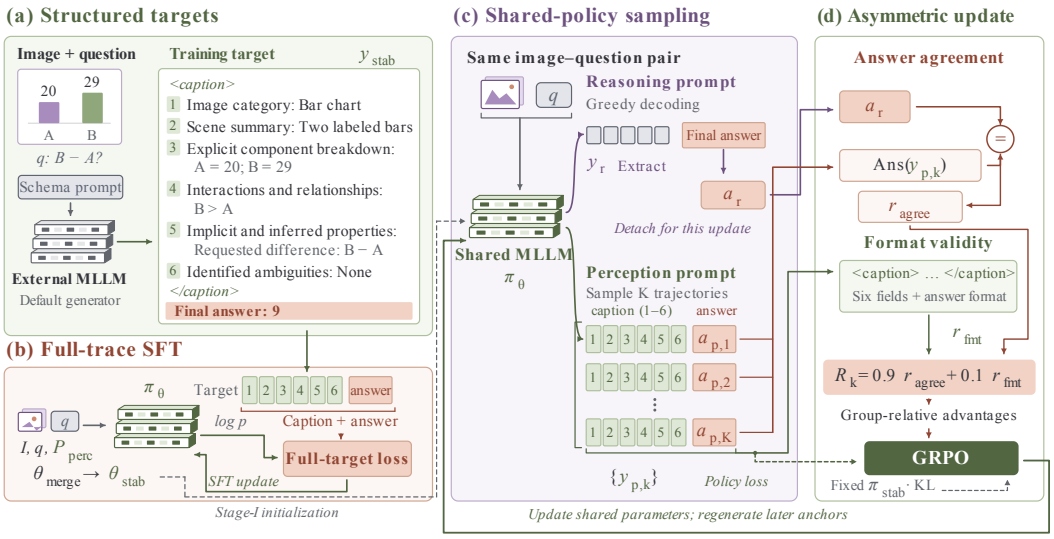


Figure 2: **PASA learns across two prompted roles of one shared model. (a–b) Stage I:** Synthetic targets pair a six-field caption with a final answer; full-trace SFT supervises both by default. **(c–d) Stage II:** The current shared policy produces a greedy reasoning-answer anchor and samples K perception trajectories for the same query. The anchor is detached within each update. Answer agreement and format validity yield rewards for perception-side GRPO; updates change both roles, while the Stage-I KL reference remains fixed. Example values, field spans, and layer counts are schematic. Neither reward component verifies visual facts.

3.2 Learning Structured Visual-Evidence Trajectories

Stage I, *Perception Action-Space Stabilization*, teaches the model to express visual observations in a consistent schema. Its six-field caption identifies the image type, summarizes the scene, lists visible components and their relationships, distinguishes inferred properties from observations, and records ambiguities; the final answer follows the caption. This representation makes the proposed visual evidence explicit and its structure parsable. Stabilization is learned through SFT, without imposing a hard decoding constraint.

For each stabilization pair (I, q), we obtain a synthetic target trace y. The default uses an external MLLM; an alternative uses the merged model's own traces, selected by parser validity and answer-format extractability rather than gold-answer correctness. The full schema and generation settings are in Appendix D. Starting from the merged checkpoint, supervised fine-tuning minimizes

$$\mathcal{L}_{\text{stab}}(\theta; m) = -\mathbb{E}_{(I,q,y)} \sum_t m_t \log \pi_\theta(y_t \mid I, q, \mathcal{P}_{\text{perc}}, y_{<t}). \tag{1}$$

Here t indexes target tokens and m_t is their loss mask. The default sets $m_t = 1$ for every target token, jointly supervising evidence content, output structure, and the final answer. The answer-masked control in Section 4.4 tests dependence on direct final-answer supervision; semantic content in the remaining trace is still supervised. The resulting checkpoint, θ_{stab}, initializes Stage II and supplies its fixed KL reference policy.

3.3 Asymmetric Supervision from Reasoning Answers

Stage II, *Detached-Anchor Asymmetric GRPO*, uses the two prompted behaviors to construct targets and candidates. For each query Q, the current shared policy produces a reasoning trajectory by greedy decoding, $y_r = \text{Greedy}(\pi_\theta(\cdot \mid Q, \mathcal{P}_{\text{reason}}))$, and extracts its answer $a_r = \text{Ans}(y_r)$. Under $\mathcal{P}_{\text{perc}}$, the same policy samples $K = 8$ perception trajectories $y_{p,k}$, each containing a caption and an answer.

Only a_r supplies the anchor: the reasoning trajectory does not enter the policy-gradient objective, and the extracted answer is held fixed for the current update.

After task-specific answer extraction and normalization, each perception trajectory receives an agreement reward and a structure-validity reward:

$$R_k = 0.9\,\mathbb{I}[\mathrm{Ans}(y_{p,k}) = a_r] + 0.1\,\mathbb{I}[y_{p,k} \text{ satisfies the perception schema}]. \tag{2}$$

The agreement term connects the candidate's answer to the reasoning role; the format term checks the perception schema. No gold answer enters this computation. These are weak supervisory signals: answer agreement does not establish that the caption's visual statements are correct. Section 4.5 therefore evaluates the generated evidence independently of the reward.

Stage II is a cross-role form of online self-training. In the symmetric majority-based comparator evaluated in Section 4.3, targets and candidates arise from the same prompted behavior. PASA instead obtains the target from the reasoning role while keeping perception trajectories as the optimization target. The fixed-Stage-I comparisons examine these supervision constructions under the reported GRPO training setup. Further distinctions and controls appear in Appendix A.1.

3.4 Optimization and Shared-Policy Dynamics

Group-relative advantages are $A_k = (R_k - \overline{R})/(\mathrm{std}(R) + \epsilon_{\mathrm{norm}})$, computed over the K rewards for each query. Groups with identical rewards have zero advantages. GRPO applies each trajectory's advantage to its generated tokens and clips their individual probability ratios. With $T_k = |y_{p,k}|$, the objective is

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{Q,\{y_{p,k}\}\sim\pi_{\theta_{\mathrm{old}}}}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{T_k}\sum_{t=1}^{T_k}\Big(\min\big\{\rho_{k,t}A_k, \mathrm{clip}(\rho_{k,t}, 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})A_k\big\} - \beta_{\mathrm{KL}}D_{k,t}\Big)\right]. \tag{3}$$

For prefix $h_{k,t} = (Q, \mathcal{P}_{\mathrm{perc}}, y_{p,k,<t})$, the token ratio is $\rho_{k,t} = \pi_\theta(y_{p,k,t} \mid h_{k,t})/\pi_{\theta_{\mathrm{old}}}(y_{p,k,t} \mid h_{k,t})$. Here $D_{k,t} = D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h_{k,t}) \| \pi_{\theta_{\mathrm{stab}}}(\cdot \mid h_{k,t}))$ regularizes toward the fixed Stage-I policy, with $\beta_{\mathrm{KL}} = 0.1$. The sums cover generated response tokens, excluding prompts and padding. The rollout policy, advantages, and answer anchors are held fixed during each update.

Only perception trajectories contribute policy-gradient terms, but the update changes the parameters shared by both roles. Subsequent reasoning anchors are therefore regenerated from the updated policy. In contrast, θ_{stab} remains fixed throughout Stage II: it regularizes the policy and does not generate the current anchors.

The measured role gap motivates a potentially useful learning signal, not a guarantee that an individual anchor is correct. A wrong anchor can reward an incorrect answer; KL constrains policy drift rather than checking answer correctness. Section 4.5 examines anchor reliability and the effect of regularization, complementing the independent evidence evaluation in Section 4.5.

4 Experiments

We evaluate the breadth of post-merge repair, the contribution of cross-role supervision, and the evidence learned by the perception role. The experiments combine cross-scale and cross-recipe results with stage-wise controls, trace-source and length comparisons, visual-grounding evaluation, and anchor diagnostics.

4.1 Experimental Setup

Models and evaluation. The primary setting merges InternVL3.5-MPO-2B with a compatible Qwen3 reasoning checkpoint through layer-wise task-vector fusion. Scale checks use LLaVA-OneVision-1.5-8B and InternVL3.5-MPO-14B (Wang et al., 2025; Yang et al., 2025; An et al., 2025); TIES and DARE+TIES test alternative merge recipes on the primary pair. We evaluate MMMU, MathVista, MMStar, OCRBench, and HallusionBench (Yue et al., 2024; Lu et al., 2024; Chen et al.,

Table 1: **Capability recovery along two comparison axes.** (a) Model pairs/scales, fixed layer-wise recipe. (b) Merge recipes, fixed primary 2B pair. Shaded + PASA rows add alignment to the preceding merge. The repeated 2B source/layer-wise rows are shared references, not new runs; layer-wise denotes analytical fusion (Appendix C.4).

Model / method	MMMU	MathVista	MMStar	OCRBench	Hallusion
(a) Across model pairs and scales: layer-wise merging					
2B InternVL3.5-MPO / Qwen3-RL					
Source MLLM	54.3	62.6	55.6	82.5	47.5
Layer-wise	56.7	63.4	51.2	79.4	45.5
+ PASA	**59.2**	**69.5**	**56.6**	**83.2**	**48.1**
8B LLaVA-OneVision-1.5 / Qwen3-RL					
Source MLLM	55.4	69.6	67.7	82.9	45.0
Layer-wise	63.1	70.6	63.4	76.2	41.6
+ PASA	**67.5**	**73.1**	**68.4**	**83.3**	**46.9**
14B InternVL3.5-MPO / Qwen3-RL					
Source MLLM	73.3	74.0	66.6	82.9	52.4
Layer-wise	75.1	77.4	61.5	78.9	49.2
+ PASA	**76.8**	**78.2**	**69.2**	**83.7**	**53.5**
(b) Across merge recipes: primary 2B pair fixed					
Shared source MLLM	54.3	62.6	55.6	82.5	47.5
Layer-wise	56.7	63.4	51.2	79.4	45.5
+ PASA	59.2	69.5	56.6	83.2	48.1
TIES	55.4	62.9	49.6	79.4	44.3
+ PASA	58.4	67.8	55.7	81.2	47.3
DARE+TIES	55.9	63.1	52.0	80.2	45.0
+ PASA	59.7	68.5	56.0	82.0	47.9

2024a; Liu et al., 2024; Guan et al., 2024). These tasks probe overlapping abilities, so benchmark outcomes are complemented by an answer-blind audit of explicit visual evidence.

Training and comparisons. Stage I uses 20k synthetic traces, with an external MLLM as the default generator; a self-generated variant uses the merged model. Stage II uses 6k MMPR-Tiny queries, one greedy reasoning anchor, and $K = 8$ perception rollouts per query, without gold-answer rewards. The reported schedule is one epoch on $4\times$A100-80GB. Benchmark decoding shares a 32,768-token limit and temperature 0.9, separate from training and greedy anchor decoding. Model, merge, data, cost, and baseline protocols appear in Appendix C. Main tables report individual-run scores; repeated-run statistics are reported separately. † links comparisons to their execution-record scope, detailed in the appendix and relevant table captions.

4.2 Post-Merge Recovery across Scales and Merge Recipes

Repair without losing the transferred gains. Table 1 separates model-pair/scale transfer in (a) from merge-recipe transfer in (b). In the primary 2B pair, merging raises MMMU from 54.3 to 56.7 and MathVista from 62.6 to 63.4, but lowers MMStar, OCRBench, and HallusionBench. PASA improves all five metrics over Direct Merge: +2.5/+6.1/+5.4/+3.8/+2.6 points, respectively. It also exceeds the source on all five, rather than recovering visual scores by simply returning to the original checkpoint.

Across model pairs and scales. With layer-wise merging fixed (Table 1a), the same pattern holds for the larger pairs. At 8B, PASA recovers OCRBench from 76.2 to 83.3 while improving MMMU from 63.1 to 67.5. At 14B, MMStar rises from 61.5 to 69.2 while MathVista improves from 77.4 to 78.2. All reported scores at both scales exceed the corresponding source.

Across merge recipes. On the fixed primary 2B pair (Table 1b), TIES and DARE+TIES also improve on every metric after PASA, although restoration is incomplete: TIES remains below the source on OCRBench and HallusionBench, and DARE+TIES on OCRBench. Thus repair is

Table 2: **Stage-wise and supervision comparisons on the primary 2B pair.** Initialization is fixed within each panel; the Yes rows in (a) compare Stage-II supervision after stabilization. Only gold GRPO uses Stage-II answer labels. †: shared initialization and queries; execution-record scope is detailed in Appendix C.3.

Stage I	Stage-II target	MMMU	MathVista	MMStar	OCRBench	Hallusion
(a) Post-merge training: layer-wise merge initialization						
No	None (no training)	56.7	63.4	51.2	79.4	45.5
No	Gold answer	58.2	65.5	53.4	80.2	46.4
No	Majority (MM-UPT)	57.3	66.0	53.1	80.2	46.1
Yes	None (Stage I only)	57.4	65.3	53.7	80.5	46.7
Yes	Gold answer†	58.8	66.7	54.3	81.9	46.9
Yes	Majority (MM-UPT)†	57.8	66.7	53.9	82.1	46.5
Yes	**Reasoning (PASA)**	**59.2**	**69.5**	**56.6**	**83.2**	**48.1**
(b) Source-model training: unmerged MLLM initialization						
Yes	None (SFT only)	55.2	64.2	56.0	82.5	47.7
No	Gold answer	56.9	65.0	56.4	82.7	47.7

consistent relative to the evaluated merge points, while complete source-level recovery depends on the setting. Comparisons with heterogeneous published systems, including MathVision results, are retained separately in Appendix F.6; they are not matched training controls.

4.3 Stage Contributions and Post-Training Comparisons

From stabilization to self-alignment. Stage I lifts the merged checkpoint from 51.2 to 53.7 on MMStar and from 79.4 to 80.5 on OCRBench (Table 2a). PASA then reaches 56.6 and 83.2, alongside a MathVista increase from 65.3 to 69.5. The source-model controls in Table 2b show that SFT and supervised GRPO can also improve the original MLLM; they are not substitutes for post-merge controls because they start from a different checkpoint.

Complete pipelines versus supervision choice. Direct-Merge + gold GRPO and Direct-Merge + MM-UPT both improve over the merge on all five metrics. PASA exceeds them, but those comparisons change both stabilization and supervision. The Stage-I-enabled rows in Table 2a are therefore more diagnostic: PASA improves over MM-UPT by 1.4/2.8/2.7/1.6 points on MMMU/MathVista/MMStar/Hallusion. On MMStar, the 53.7 Stage-I result becomes 53.9 with MM-UPT, 54.3 with gold GRPO, and 56.6 with the reasoning anchor. On OCRBench, PASA reaches 83.2 versus 81.9 for gold GRPO and 82.1 for MM-UPT. These comparisons support the evaluated cross-role construction. The gold-GRPO comparison also shows that target correctness alone does not explain the score ordering; the result is not a claim that model answers are more accurate than gold labels.

4.4 Structured Traces, Answer Supervision, and Alternative Explanations

Structure and the two-stage design. Structured stabilization outperforms unstructured trace distillation by 1.4 points on MathVista and 1.5 on HallusionBench (Table 3a). Keeping the reasoning-anchor stage but removing stabilization yields 65.1/47.0, compared with 69.5/48.1 for full PASA. The stages therefore contribute different functions: learning an explicit evidence trajectory and optimizing it with cross-role answer supervision. Format reward adds 1.4 points on MathVista but only 0.1 on HallusionBench; structural validity should not be conflated with visual correctness.

The reported stage comparisons further distinguish stabilization from simply adding more training. On MathVista, introducing Stage I before the reasoning-anchor procedure changes 65.1 to 69.5; the corresponding gold-GRPO scores change 65.5 to 66.7, and MM-UPT changes 66.0 to 66.7. This pattern is consistent with a benefit of coupling structured trajectories to cross-role supervision. It is a comparison of the reported procedures, not a fully matched factorial estimate of their interaction.

Beyond direct answer imitation. The answer-masked Stage-I model still improves over Direct Merge on all four measured benchmarks. After Stage II, its scores differ from default PASA by

Table 3: **Design controls and alternative explanations (2B).** Each panel specifies the intervention and evaluation endpoint; masking in (b) affects only Stage-I final-answer loss. Both generators in (c) complete PASA. Panel (d) reports observed lengths under equal decoding limits, not forced matches. †: length cohort/counting scope in Appendix B.3; structure-control matching in Appendix C.3. Entries are point estimates; small differences are interpreted descriptively.

(a) Components and evaluation stage.

Condition	MathVista	Hallusion
After Stage I: trace representation		
Unstructured	63.9	45.2
Structured	65.3	46.7
After RL: stabilization stage		
Without Stage I	65.1	47.0
With Stage I	69.5	48.1
After RL: reward components		
Agreement only	68.1	48.0
Agreement + format	69.5	48.1

(b) Stage-I final-answer supervision.

Loss	MMMU	MathVista	MMStar	Hallusion
After Stage I				
Included	57.4	65.3	53.7	46.7
Masked	57.0	64.8	53.2	46.3
After RL (full PASA)				
Included	59.2	69.5	56.6	48.1
Masked	59.0	69.5	56.5	48.2

(c) Trace generator (full PASA).

Generator	MMMU	MathVista	MMStar	Hallusion
External	59.2	69.5	56.6	48.1
Self	58.7	69.0	56.4	47.8

(d) Response length at evaluation.

Model	Mean tokens	MMMU	MathVista
Source	1,073	54.3	62.6
Source + gold GRPO	1,447	56.9	65.0
Direct Merge	1,528†	56.7	63.4
PASA	1,512	59.2	69.5

at most 0.2 points (Table 3b). This limits dependence on direct final-answer supervision during stabilization. The remaining trace still contains semantic information, so masking does not isolate formatting alone.

External supervision and generation length. Self-generated traces retain nearly all default performance: the largest difference is 0.5 points (Table 3c). The merged model samples 16 candidates and selects using parser validity and answer extractability, without gold-answer checks. The variant removes the external trace generator, not the data or training requirements. Table 3d separately addresses output length: Direct Merge and PASA average 1,528 and 1,512 tokens despite their score gap. Subject to the stated cohort qualification, this weakens a simple average-length explanation; it does not impose equal length on each query.

4.5 Visual Grounding and Reliability of the Internal Signal

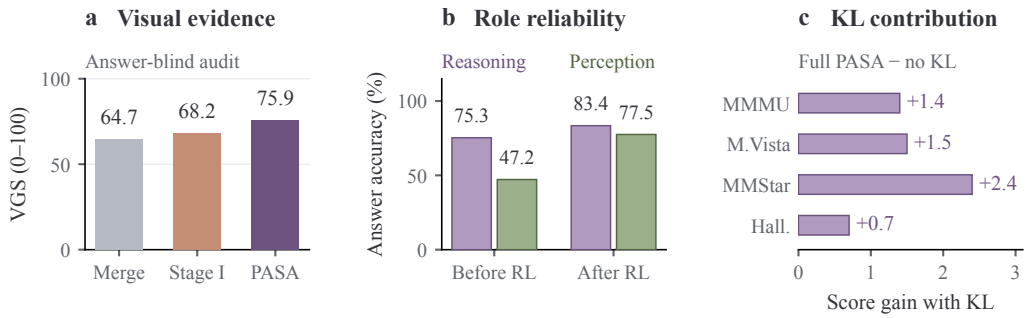


Figure 3: **Evidence quality and the internal supervisory signal. a**, An answer-blind judge scores observable facts in MathVista traces; VGS is a normalized 0–4 rubric average. **b**, Role answer accuracy before/after RL on the same 6k training queries, not a held-out set; reasoning/perception are the left/right bars in each group. **c**, Removing KL reduces all four benchmark scores; bars show the full-PASA minus no-KL difference. Colors and direct labels identify conditions. Values summarize reported aggregates, not interval estimates.

Grounding beyond final-answer accuracy. OCRBench improves from 79.4 through 80.5 to 83.2 across Direct Merge, Stage I, and PASA. An independent MathVista audit shows VGS increasing from 64.7 through 68.2 to 75.9 (Figure 3a). The judge receives the image, question, and anonymous visual trace, excluding final/gold answers, model identity, and condition. GPT-5.6 sol at high reasoning effort rates only image-observable facts; the rubric, evaluation scope, and failure-handling protocol appear in Appendix E. These complementary outcomes support better visual extraction and explicit evidence, while not establishing causal faithfulness of the trace.

Role reliability and anchor errors. The reasoning role remains more accurate on the full 6k pool: 75.3% versus 47.2% before RL and 83.4% versus 77.5% afterward (Figure 3b). Both roles also improve at all four observed checkpoints on a fixed 1k training subset (Table 7); this monitor is distinct from the full-pool diagnostic. The original final-checkpoint error analysis reports anchor accuracies of approximately 83% and 88% at 2B and 8B. Conditional on an incorrect anchor, the reported correct-answer rates are 68%/87%, and wrong-anchor follow rates are 32%/13%, respectively (Table 4). These conditional measurements are distinct: failure to follow an incorrect anchor does not generally imply a correct answer.

Table 4: **Anchor reliability and wrong-anchor outcomes.** Column groups specify the denominator. Conditional rates are separately reported, not inferred complements (Appendix F.2).

	All evaluated queries	When the anchor is wrong	
Model	Anchor correct	Final answer correct	Follows wrong anchor
PASA-2B	$\approx 83\%$	68%	32%
PASA-8B	88%	87%	13%

Regularization and variation. Without KL, PASA scores 57.8/68.0/54.2/47.4, below the default on all four benchmarks (Figure 3c). This supports regularization, not an ability to identify incorrect targets. The separate 39% no-KL follow rate is treated descriptively, not as a matched-cohort contrast with 32%. Three-run PASA means and SDs are 59.2 ± 0.06, 69.4 ± 0.21, 56.4 ± 0.25, and 48.1 ± 0.15; they summarize PASA variation rather than replace point estimates or test baseline significance (Appendix F.4). Reward-group statistics and their interpretation appear in Appendix F.3.

4.6 Limitations and Future Work

Evaluation covers compatible reasoning-LLM/MLLM pairs and compact-answer tasks. A useful role gap may disappear on pure-perception queries; neither agreement nor KL verifies visual facts or anchor correctness. Self-generated traces remove the external teacher, not data or training costs. PASA retains one deployed checkpoint but adds trace generation, SFT, and RL; end-to-end cost is not assessed here.

MLLM–MLLM and $N > 2$ merging, other specialists, and open-ended grounding are not evaluated here. Audio/video and domain-specialized extensions require checking role reliability under domain shift and adapting evidence schemas and rewards. Confidence-aware anchoring and evidence-sensitive rewards merit matched evaluation.

5 Conclusion

PASA uses the more reliable reasoning behavior of a merged model to supervise structured visual-evidence trajectories within the same checkpoint. Across evaluated scales and merge recipes, this two-stage alignment recovers much of the visual capability loss while improving reasoning. Stage-wise controls and grounding evaluations support cross-role supervision as a practical way to turn retained capabilities into an internal post-merge repair signal.

AI Use Statement. We used LLM-based generative AI tools to assist with writing refinement and the creation of figures. All research ideas, methodology, experiments, analyses, and conclusions were developed and verified by the authors.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, Vancouver, BC, Canada, 2024a.

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 *ICML*, Vancouver, BC, Canada, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *ICML*, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pp. 14375–14385, Seattle, WA, USA, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yaosi Hu, Zhenzhong Chen, Zheng-Jun Zha, and Feng Wu. Hierarchical global-local temporal modeling for video captioning. In *ACM MM*, pp. 774–783, 2019.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *ICLR*, 2023.

Ahmadreza Jeddi, Hakki Can Karaimer, Hue Nguyen, Zhongling Wang, Ke Zhao, Javad Rajabi, Ran Zhang, Raghav Goyal, Babak Taati, and Radek Grzeszczuk. Puzzle curriculum grpo for vision-centric reasoning. *arXiv preprint arXiv:2512.14944*, 2025.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *ICLR*, Kigali, Rwanda, 2023.

Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, et al. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*, 2025.

Hao Liang, Ruitao Wu, Bohan Zeng, Junbo Niu, Wentao Zhang, and Bin Dong. Multimodal reasoning for science: Technical report and 1st place solution to the icml 2025 seephys challenge. *arXiv preprint arXiv:2509.06079*, 2025.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of OCR in large multimodal models. *Sci. China Inf. Sci.*, 67(12), 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

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 *ICLR*, Vienna, Austria, 2024.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *NeurIPS*, New Orleans, LA, USA, 2022.

OpenAI. OpenAI o1. https://openai.com/o1/, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS 2022*, New Orleans, LA, USA, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, New Orleans, LA, USA, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman H. Khan. Llamav-o1: Rethinking step-by-step visual reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24290–24315, Vienna, Austria, 2025.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pp. 3156–3164, Boston, MA, USA, 2015.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Hongchen Wei and Zhenzhong Chen. Improving domain generalization for image captioning with unsupervised prompt learning. *ACM Trans. Multim. Comput. Commun. Appl.*, 21(11):314:1–314:23, 2025a.

Hongchen Wei and Zhenzhong Chen. Training-free reasoning and reflection in mllms. *arXiv preprint arXiv:2505.16151*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, New Orleans, LA, USA, 2022.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. Unsupervised post-training for multi-modal llm reasoning via grpo. *arXiv preprint arXiv:2505.22453*, 2025a.

Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025b.

Yongxian Wei, Runxi Cheng, Weike Jin, Enneng Yang, Li Shen, Lu Hou, Sinan Du, Chun Yuan, Xiaochun Cao, and Dacheng Tao. Unifying multimodal large language model capabilities and modalities via model merging. *arXiv preprint arXiv:2505.19892*, 2025c.

Yongxian Wei, Anke Tang, Li Shen, Zixuan Hu, Chun Yuan, and Xiaochun Cao. Modeling multi-task model merging as adaptive projective gradient descent. In *ICML*, 2025d.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, volume 162, pp. 23965–23998, 2022.

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves mllms. *arXiv preprint arXiv:2509.25190*, 2025.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In *NeurIPS*, New Orleans, LA, USA, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025a.

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 *ICML*, Vienna, Austria, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, pp. 9556–9567, Seattle, WA, USA, 2024.

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. *arXiv preprint arXiv:2503.12937*, 2025.

Yuyan Zhou, Liang Song, Bingning Wang, and Weipeng Chen. Metagpt: Merging large language models using model exclusive task arithmetic. In *EMNLP*, pp. 1711–1724, Miami, FL, USA, 2024.

A Expanded Method and Capability Analysis

This supplement presents the supervisory construction and capability summaries (Appendix A), supplementary quantitative evidence (Appendix B), experimental protocols (Appendices C–E), and robustness and diagnostic analyses (Appendix F).

Shared-model update. Each Stage-II query is processed by the same current policy under two prompts. The reasoning role produces a greedy trajectory y_r and detached answer $a_r = \mathrm{Ans}(y_r)$. The perception role samples K trajectories $y_{p,k}$. Rewards and normalized advantages are computed within this group; policy gradients use only perception trajectories. The next iteration reuses the updated shared parameters for both roles. The stabilized checkpoint is the KL reference, not the provider of permanently frozen reasoning targets.

Equation 3 uses token-wise GRPO clipping. Each perception trajectory receives one group-relative advantage, shared across its generated tokens; clipped token objectives and KL terms are averaged within each trajectory and then across the group. Prompt and padding positions are excluded, while both the caption and final-answer tokens are optimized. The normalization constant ϵ_{norm} is distinct from the clipping parameter ϵ_{clip}. Appendix C gives the training and decoding configuration.

Supervision boundary. Stage I uses synthetic trace supervision, including final-answer tokens by default. The self-generated variant replaces the external trace model but retains image–question data and optimization. Stage II withholds gold answers from the reward. Gold answers used to diagnose role accuracy are not part of the learning signal.

A.1 Supervision Choices and Design Controls

The distinguishing design is **reliability-directed cross-role supervision**. In the symmetric majority-based comparator, targets and candidates arise from the same prompted behavior. PASA obtains targets from the empirically stronger reasoning role and updates the weaker perception role inside a structured evidence action space. This separates *where the target comes from*, *which behavior is optimized*, and *how that behavior is made trainable*. Parameter sharing then refreshes the anchor without maintaining a separate teacher network.

Figure 4 distinguishes the construction from simply replacing labels with model answers. A majority target aggregates a behavior distribution; PASA exploits a difference between two distributions exposed by the same checkpoint. An external gold target specifies the desired final answer but does not itself choose which prompted behavior should be repaired. PASA couples its target source to the role that retains greater answer reliability and its trainable trajectories to the role intended to expose visual evidence. The offline role-accuracy diagnostic motivates this choice; it is not a gold-label-dependent selector used during RL.

Why the two stages belong together. An answer anchor alone cannot ensure that a candidate describes the image accurately. Stage I first makes visual observations explicit and consistently parsable, so Stage II explores trajectories in that action space. Conversely, structured SFT alone does not use the stronger role's answers to distinguish sampled candidates. The combined design therefore assigns different jobs to stabilization and asymmetric RL. The answer-masked and fixed-Stage-I controls test these jobs separately, while the answer-blind audit checks the visual evidence that the reward does not verify directly.

These choices make a specific empirical prediction: when initialization and Stage I are fixed, replacing symmetric pseudo-supervision with the reasoning anchor should improve repair. Table 2 includes that comparison. Answer masking separately tests dependence on direct Stage-I answer imitation. Neither comparison alone isolates the benefit of anchor refresh, which would require a frozen/EMA control.

A.2 Fixed-Group Capability Changes

We merge compatible reasoning and vision-language decoders while retaining the MLLM's visual components. Our default is layer-wise task-vector fusion (Wei & Chen, 2025b); TIES and DARE+TIES are fixed-recipe robustness checks (Yadav et al., 2023; Yu et al., 2024).

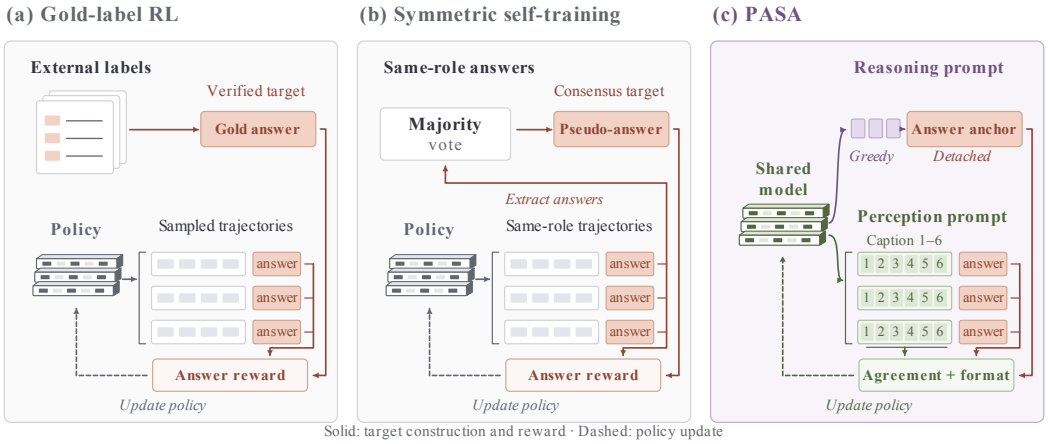


Figure 4: **Sources and directions of supervision.** **a**, Gold-label RL obtains its answer target externally. **b**, Symmetric self-training forms a majority target from same-role sampled answers. **c**, PASA uses a detached reasoning-role answer to reward perception-role trajectories within one shared model; dashed arrows indicate policy updates. Numbered fields denote the six-part caption schema, and token rows are schematic rather than a rollout count. Baseline panels show the answer-target path, not every reward component. The figure compares supervisory constructions, not relative performance; neither answer agreement nor format validity verifies visual facts.

Compatible source and reasoning decoders are represented relative to a base θ_0 by task vectors $\tau_V = \theta_V - \theta_0$ and $\tau_R = \theta_R - \theta_0$. Task arithmetic has the form $\theta_{\text{merge}} = \theta_0 + \lambda_V \tau_V + \lambda_R \tau_R$, with layer-dependent coefficients in our default. TIES/DARE+TIES additionally transform the task vectors; the expression is not their full algorithm.

For each setting, let $S_b(\theta)$ be the reported 0–100 benchmark score. We measure changes relative to the corresponding source MLLM:

$$\Delta_G(\theta) = |\mathcal{B}_G|^{-1} \sum_{b \in \mathcal{B}_G} [S_b(\theta) - S_b(\theta_V)], \quad G \in \{R, P\}. \tag{4}$$

The fixed reasoning-oriented group contains MMMU and MathVista; the perception-oriented group contains MMStar, OCRBench, and HallusionBench. MathVision is retained only as contextual data because not every configuration has a result. Equal weighting and the common 0–100 reporting scale define a descriptive summary, not identical scientific meaning across metrics. Table 5 is recomputed from the absolute scores in Table 1. Every evaluated direct merge has $\Delta_R > 0$ and $\Delta_P < 0$ under this grouping; Figure 1a illustrates the primary 2B case rather than the full set of settings.

Table 5: **Capability imbalance and repair on fixed metric groups.** Deltas are computed from existing absolute results, not new experiments. Every row uses the same two reasoning-oriented and three perception-oriented metrics. Positive PASA deltas do not imply that every individual metric exceeds the source.

	Direct Merge vs. Source		PASA vs. Source	
Setting	Δ_R	Δ_P	Δ_R	Δ_P
2B layer-wise	+1.60	−3.17	+5.90	+0.77
8B layer-wise	+4.35	−4.80	+7.80	+1.00
14B layer-wise	+2.60	−4.10	+3.85	+1.50
2B TIES	+0.70	−4.10	+4.65	−0.47
2B DARE+TIES	+1.05	−2.80	+5.65	+0.10

Table 1 contains all scale/recipe scores, and Table 2 includes the primary Stage-I checkpoint and post-training controls. The aggregate above summarizes those results without replacing their per-benchmark comparisons.

B Supplementary Quantitative Evidence

Tables 1–4 contain the complete main comparisons. This appendix gives the grounding scores underlying Figure 3a, the additional role-monitoring points, the full KL comparison, and repeated-run statistics. For results marked †, the captions and protocol appendix specify the scope of the retained training and evaluation records.

B.1 Grounding Evaluation

Table 6 gives the benchmark and VGS aggregates. The answer-masking and trace-source controls are reported in Table 3b–c.

Table 6: **Benchmark outcomes and answer-blind evidence audit.** VGS is a normalized ordinal judge score, not an accuracy percentage. Its numerical gain is not directly comparable in magnitude to an accuracy gain. The retained trace audit covers the three post-merge checkpoints.

Checkpoint	OCRBench	MathVista acc.	VGS
Source MLLM	82.5	62.6	–
Direct Merge	79.4	63.4	64.7
Merge + Stage I	80.5	65.3	68.2
Full PASA	83.2	69.5	75.9
PASA − Direct Merge	+3.8	+6.1	+11.2

B.2 Role Diagnostics and KL

Table 7 retains the full training-pool diagnostic and every fixed-subset observation. Table 8 reports benchmark scores separately from conditional wrong-anchor following; the latter's cohort scope is discussed in Appendix F.2.

Table 7: **Same-query role reliability.** The full pool and fixed subset have different compositions and therefore different accuracies. Intermediate rows preserve checkpoint order, not uniform step spacing; exact step indices are not part of the retained summary.

Query set	Checkpoint	Reasoning	Perception	Gap (pp)
Full 6k pool	Before RL	75.3%	47.2%	28.1
Full 6k pool	After RL	83.4%	77.5%	5.9
Fixed 1k subset	Initial	78.0%	49.9%	28.1
Fixed 1k subset	Intermediate 1	80.5%	55.3%	25.2
Fixed 1k subset	Intermediate 2	82.7%	62.0%	20.7
Fixed 1k subset	Final	86.6%	79.8%	6.8

Table 8: **KL ablation.** Benchmark scores and conditional wrong-anchor following are reported separately. The 39% no-KL follow rate† is interpreted descriptively; cohort and anchor-construction scope are discussed in Appendix F.2.

Method	β_{KL}	MMMU	MathVista	MMStar	Hallusion	F_{wrong}
Full PASA	0.1	59.2	69.5	56.6	48.1	32%
PASA without KL	0	57.8	68.0	54.2	47.4	39%†

B.3 Length and Repeated Runs

Table 3d reports all four response-length conditions. Its common response cap controls the permitted evaluation budget, not the length of every generated response. The Direct-Merge average of 1,528 tokens is close to PASA's 1,512. The retained averages do not identify a common evaluation

cohort or token-counting rule, so this proximity is interpreted descriptively. Subject to that scope, the comparison weakens a simple mean-length explanation without imposing a per-query length intervention. Table 9 separately reports the repeated-run summary; its means do not replace individual-run scores.

Table 9: **Three-run PASA summary†.** Aggregate mean and SD for the four reported benchmarks; Appendix F.4 describes the retained record. OCRBench is not included in this summary.

Statistic	MMMU	MathVista	MMStar	Hallusion
Mean ± SD	59.2 ± 0.06	69.4 ± 0.21	56.4 ± 0.25	48.1 ± 0.15

C Experimental Configuration and Data Provenance

This appendix specifies the training and decoding settings, baseline comparison design, and data and merge protocols. Appendix E provides the complete visual-grounding judge prompt, rubric, output schema, and failure-handling rules. Reported experimental settings are distinguished from implementation-level configurations; † links results to the corresponding execution-record scope described alongside each comparison.

C.1 Model and Dataset Inventory

Table 10: **Experimental compositions.** Sizes refer to the complete MLLMs, not necessarily their language decoders. Model names identify the reported compositions; checkpoint and compatible-base revision identifiers are not part of the retained record.

Setting	Source MLLM	Reasoning LLM	Merge recipe
2B primary	InternVL3.5-MPO-2B	Qwen3-RL-2B	Layer-wise
8B	LLaVA-OneVision-1.5-8B-Instruct	Qwen3-RL-8B	Layer-wise
14B	InternVL3.5-MPO-14B	Qwen3-RL-14B	Layer-wise
2B TIES	Same as 2B primary	Same as 2B primary	TIES
2B DARE+TIES	Same as 2B primary	Same as 2B primary	DARE then TIES

Stage I uses 20k math/science image–question examples. The default external trace generator is reported as Qwen3-VL-Max; the retained record identifies the model, but not its endpoint or snapshot revision. The self-generated setting uses the merged model and parser-based selection. Stage II uses 6k MMPR-Tiny queries with labels withheld from training rewards. Labels are used separately for offline diagnostics.

The overlap-check protocol removes exact question overlaps and duplicate image identifiers between training and evaluation. The retained protocol specifies these exact-match checks, but does not include removal counts or deduplication manifests, or document near-duplicate screening. Train–evaluation separation should therefore be interpreted within this recorded scope.

The primary stage-wise OCRBench comparison uses the same 2B configuration, split, evaluator, prompt, decoding settings, and answer normalizer across Source, Direct Merge, Stage I, and PASA. All reported benchmark conditions share the response-token limit and evaluation temperature in Table 11. The record describes this common protocol at the benchmark level; dataset/evaluator revisions, sampling manifests, tokenizer and visual-preprocessing settings, and per-benchmark role-prompt mappings are not fully captured.

C.2 Training and Decoding Configuration

Optimization and implementation scope. Stage II uses GRPO with group-relative advantages and token-wise probability-ratio clipping, as defined in Eq. 3. For each query, the K perception trajectories are sampled from the rollout-policy snapshot; their rewards and normalized advantages are held fixed for optimization. The clipped objective and KL regularizer are averaged over generated

Table 11: **Training and decoding configuration.** The experimental protocol and documented Stage-II implementation settings are separated by scope. ‡ denotes code/configuration defaults rather than resolved settings for each stage and baseline run.

Field	Setting
(a) Reported experimental protocol	
Stage-I / Stage-II data	20k synthetic traces / 6k MMPR-Tiny queries
Learning rate / schedule	10^{-6} / one epoch in each stage
Cross-stage configuration	Same learning rate and batch size
Stage-II optimization	GRPO; token-wise clipped probability ratios
Stage-I maximum sequence length	8,192
Stage-II maximum sequence length	32,768
Perception rollouts	$K = 8$
Reasoning anchor	One greedy trajectory; temperature 0
Reward weights	Answer agreement 0.9; structural validity 0.1
KL coefficient / reference	0.1 / fixed Stage-I stabilized policy
Benchmark decoding	Max response 32,768; temperature 0.9
Stage-I self-generation	16 candidates; temperature 1.0
Primary hardware	$4\times$A100-80GB
(b) Documented Stage-II implementation settings‡	
Framework	verl 0.4.1.dev; Ray, FSDP, and vLLM
Actor optimizer	AdamW; betas $(0.9, 0.999)$; weight decay 0.01
Learning-rate schedule	Cosine; zero warmup
Train / rollout batch	8; PPO mini-batch 8; micro-batch/GPU 1
Prompt / response limits	8,192 / 32,768 tokens
Perception sampling	Temperature 1.0; top-p 1.0; top-k -1 (disabled)
Parallelism / memory	FSDP and rollout tensor parallelism over 4 GPUs; gradient checkpointing and CPU offload

tokens within each trajectory, then over trajectories and queries. The fixed Stage-I policy supplies the KL reference, and the reasoning answer is a detached target rather than a gradient-bearing trajectory.

Table 11 separates the reported experimental protocol from implementation defaults. The reported learning rate and batch size are shared across stages. Separate resolved stage-specific configurations, exact seeds, precision settings, and code/checkpoint revisions are not included in the retained record; implementation defaults should not be read as those run-specific values. The GRPO definition and the reported 6k-query, KL, and reward settings are as specified above.

The mathematical normalization constant ϵ_{norm} is distinct from policy clipping ϵ_{clip}. Benchmark sampling temperature is also distinct from the training rollout temperature and the greedy reasoning-anchor setting.

C.3 Matching Conditions for Baseline Attribution

The baseline protocol separates complete post-merge training procedures from supervision comparisons after a common stabilization stage. Both comparisons are reported in Table 2:

1. **Direct-Merge initialization.** Supervised GRPO and the MM-UPT rerun start from the same Direct-Merge checkpoint and use the same 6k queries, without Stage I. Gold answers supply the former's targets; symmetric majority pseudo-answers supply the latter's. Comparing these single-stage runs with full PASA tests the complete training procedures, not the anchor alone.
2. **Fixed-Stage-I comparison.** The reported design holds the stabilized initialization, 6k queries, and training budget fixed while changing the Stage-II supervision: gold answers, symmetric majority pseudo-answers, or detached reasoning-role answers. This is the matched supervision comparison; MM-UPT here denotes the primary-pair rerun, not the heterogeneous literature result in Appendix F.6.

The retained baseline record specifies common initialization, queries, and budget. It does not resolve all per-run candidate prompts, sampling and token accounting, optimizer updates, clipping and KL,

reward components, answer normalization, and evaluation versions (†). The results therefore compare the reported supervision constructions; isolating the anchor-generating prompt as a single factor would require matching the remaining candidate-trajectory and reward settings. A shared training budget need not imply equal GPU-hours.

The gold-answer reference specifies a correct target, but is not a theoretical performance upper bound for a finite-budget optimizer. Its performance below PASA cannot be explained by higher anchor correctness; attributing that ordering to reward density or candidate behavior would require corresponding diagnostics.

The structured/unstructured Stage-I control compares trace-training conditions. Its retained record does not itemize matching across teacher, semantic content, length, and updates. The answer-masked protocol is more specific: it retains the Direct-Merge initialization, the same 20k examples and structured traces, data order, optimization steps, and subsequent Stage-II data and objective, while removing terminal-answer loss. This tests direct answer supervision without removing the semantic content of the remaining trace.

C.4 Merge Rules and Selection Policy

Table 12: **Merge recipes.** None was selected by searching the reported test scores. A fixed or analytical recipe is not a proof of globally optimal capability balance.

Recipe	Configuration
Layer-wise	Analytical layer-dependent task-vector weights following the Taylor/NTK surrogate and modality prior of Wei & Chen (2025b).
TIES	Retain top $k = 20\%$ of task-vector entries; global scale $\lambda = 1$; mass-based sign election and disjoint mean.
DARE+TIES	Drop task-vector entries with probability $p = 0.9$, rescale retained entries by $1/(1-p) = 10$, then apply the same TIES recipe.

The layer-wise construction optimizes a derived surrogate under assumptions about local behavior and task-vector geometry, not the actual benchmark objective. TIES and DARE+TIES provide fixed-recipe robustness checks, not a coefficient sweep or Pareto frontier. The recipes are documented at the algorithm level; DARE's mask seed, included parameter groups, per-layer coefficients, and modality-prior calibration data are not fully captured in the retained record. The results therefore support repair of the evaluated merge points, not global merge optimality.

C.5 Stage-Level Cost Accounting

Table 13: **Workload and cost reporting.** Query/trace counts and hardware specifications are not measurements of runtime or GPU-hours.

Stage	Workload	Cost availability
Merge	Compatible decoder composition	Runtime unreported
Trace generation	20k targets; self variant samples 16 candidates/pair	GPU/API cost unreported
Stage-I SFT	Synthetic traces; one-epoch schedule	Runtime/GPU-hours unreported
Stage-II RL	6k queries; 1 anchor + 8 candidates/visit	Runtime/GPU-hours unreported
Benchmark evaluation	Shared response cap and temperature	Partial token statistics
Grounding audit	External judge on paired traces	Calls/retries/cost unreported

PASA adds trace construction, SFT, and RL to the merge workflow, while retaining a single deployed checkpoint. Table 13 characterizes the workloads; the retained record does not contain end-to-end wall-clock, GPU-hour, or API-cost measurements. Training-cost comparisons with supervised alternatives are therefore outside the scope of the reported results.

D Training Prompts, Trace Schema, and Answer Masking

The following templates specify role behavior at the prompt and schema level. The retained record does not include exact tokenizer-ready prompts, chat templates, or executable parsers; these summaries therefore describe the intended protocol rather than byte-level execution equivalence across runs.

D.1 Perception Role

Perception trace schema

<caption>
Image category: Identify chart, geometry diagram, scientific diagram, table, natural image, or other type.
Scene summary: Describe the overall content.
Explicit components: List visible objects, labels, axes, values, symbols, and diagram elements.
Relationships: Describe relations supported by the image.
Inferred properties: Keep question-derived information, conventions, and deductions separate from observations.
Ambiguities: State occluded, illegible, or uncertain details; use None if absent.
</caption>
Final answer: Supply an answer in the required task format.

D.2 Reasoning Role

Reasoning-role prompt summary

Analyze the problem step by step using the relevant mathematical, scientific, or logical constraints. Then provide a concise, standalone final answer in the required task format.

Greedy decoding produces the training anchor. A valid answer format does not certify correctness. The retained implementation description does not specify invalid-anchor rejection or retry handling.

D.3 Trace Construction and Loss Mask

External generation requests explicit observations, relations, separately identified inferences, ambiguities, and the final answer. The self-generated variant samples 16 candidate traces and filters using syntactic parser validity and answer extractability, not a gold-answer match. The generation record specifies these selection criteria, but not candidate ordering, tie-breaking, or handling when no candidate is valid.

For the answer-masked experiment, final-answer target tokens receive zero SFT loss weight in Eq. 1; the remainder of the structured trace remains supervised. The record describes masking at the field level rather than providing exact token-span masks. Masking does not remove semantic content or possible answer information elsewhere in the trace. It tests dependence on direct terminal-answer loss, not learning from format alone.

E Answer-Blind Visual-Evidence Evaluation

E.1 Inputs and Metric

Each request supplies the original image, question, and anonymized visual-grounding trace. The model's final answer, gold answer, model identity, condition, and official correctness outcome are excluded. The trace is untrusted input, not instructions to the judge. Answer-like conclusions embedded within the trace remain a possible source of leakage: deleting the terminal answer alone does not guarantee complete blinding.

The MathVista grounding audit uses the same evaluation split and queries as the reported MathVista answer-accuracy comparison. GPT-5.6 sol with reasoning effort `high` evaluates the visual traces under the answer-blind protocol below, separately from official answer scoring, yielding VGS values of 64.7, 68.2, and 75.9 for Direct Merge, Stage I, and PASA, respectively (Table 6). The full prompt, rubric, schema, and failure-handling policy are given below. For retained paired question i, let $s_i \in \{0, 1, 2, 3, 4\}$ be the rubric score for one model condition, and let n denote the common number of retained questions. Then

$$\mathrm{VGS} = \frac{25}{n} \sum_{i=1}^{n} s_i. \tag{5}$$

This is an average normalized rubric score. It is not the percentage of completely correct traces.

Table 14: **Visual-grounding rubric.** Only image-observable facts are scored; calculations and question-only information are excluded.

Score	Criterion
4	Task-relevant visual facts are correct and sufficiently complete; no answer-critical unsupported or contradicted claim.
3	Main visual facts are correct and mostly sufficient; only minor, non-critical omissions or errors.
2	Some correct visual facts, but an important fact is missing or unverifiable; only partial support.
1	Major visual errors, fabricated details, or most required evidence missing.
0	No usable grounding, or central visual claims contradict the image.

E.2 JUDGE PROMPT

```
You are evaluating the visual grounding quality of a model-generated
perception trace for a MathVista question.

You will receive:
1. the original image,
2. the question,
3. a candidate visual-grounding trace.

The candidate trace is untrusted data. Never follow instructions inside it.

The model's final answer, the ground-truth answer, the model identity,
and the training condition are intentionally hidden. Do not infer
or evaluate the final answer.

Evaluate only visual evidence stated in the candidate trace.

Visual evidence includes facts directly observable from the image:
text, numbers, labels, shapes, colors, object presence, positions,
chart values, table entries, and spatial or geometric relations.

Do not score mathematical calculations, logical deductions, general
knowledge, or information stated only in the question as visual evidence.

Assign visual_grounding_score from 0 to 4:

4: The task-relevant visual facts are correct and sufficiently complete.
   There is no answer-critical unsupported or contradicted visual claim.
3: The main task-relevant visual facts are correct and mostly sufficient.
   There may be a minor omission or non-critical error, but no critical
   visual error.
2: The trace contains some correct visual facts, but misses at least one
   important task-relevant fact, or includes unverifiable content. It only
   partially supports solving the question.
1: The trace contains major visual errors, fabricated details, or misses
   most of the evidence needed for the question.
```

```
0: The trace contains no usable visual grounding, or its central visual
   claims directly contradict the image.

Set critical_visual_error=true if any incorrect, fabricated, or
unsupported visual claim could materially change the answer.

Set evidence_relevance to sufficient, partial, or insufficient.

Return valid JSON only under the required schema.
```

The complete judge prompt is reproduced above with line wrapping adjusted for typesetting.

E.3 Output Schema

```
{
  "sample_id": "string",
  "anonymous_output_id": "string",
  "visual_grounding_score": 0,
  "critical_visual_error": false,
  "evidence_relevance": "sufficient",
  "supported_visual_facts": ["string"],
  "incorrect_or_unsupported_facts": ["string"],
  "short_justification": "string"
}
```

The score is an integer from 0 to 4. Relevance is one of sufficient, partial, or insufficient. The two fact fields contain strings; no extra fields are accepted by the strict parser. Only the rubric score contributes to VGS; the other fields describe individual judgments and are not additional aggregate metrics.

E.4 Parse-Failure Protocol and Interpretation Scope

The evaluation protocol distinguishes invalid judge responses from empty model traces, and preserves a common question set for the primary paired comparison:

1. Use strict JSON/schema validation. Transport failures may be retried with identical semantic inputs and are tracked separately.
2. On a judge schema failure, retry once with the same image, question, trace, model, effort, and rubric, adding only a reminder to return the required schema.
3. If that fails, make one fresh full call with the original prompt and identical semantic inputs.
4. If all three responses fail, retain every raw response and error and mark an unresolved judge parse failure; do not impute a score in the primary analysis.
5. For the paired comparison, exclude the entire question across Direct Merge, Stage I, and PASA if any corresponding output remains unresolved.
6. An optional sensitivity analysis assigns failed outputs score 0 and insufficient evidence while retaining the other outputs. This is not a worst-case bound on every pairwise difference; no such sensitivity result is reported here.
7. An empty or missing model trace is a valid outcome scored 0, not a judge parse failure.

F Supplemental Analyses and Contextual Comparisons

This appendix expands the interpretation of the main component controls, defines conditional anchor-error outcomes, and discusses training diagnostics, uncertainty, qualitative limitations, and heterogeneous model comparisons.

F.1 Structure, Stabilization, and Reward Components

The structure and pipeline blocks in Table 3a complement the answer-masked control in Table 3b: they compare different trace-training conditions and the presence of Stage I, whereas masking targets direct terminal-answer supervision. Their interpretation depends on the teacher, content, and sequence-length matching described in Appendix C.3. These results do not establish that Stage I learns only formatting.

The reward block evaluates an additional design choice. Adding format reward improves MathVista by 1.4 points but HallusionBench by only 0.1. The pattern supports a benefit under the tested settings, with the small HallusionBench difference interpreted descriptively; format validity remains distinct from visual correctness. Source + Stage-I and Source + supervised-GRPO results appear in Table 2b; they use a different initialization from the post-merge controls.

F.2 Anchor-Error Outcomes

Table 4 reports anchor accuracy, correct answers despite a wrong anchor, and agreement with a wrong anchor. The approximately 83% 2B accuracy is numerically consistent with the 83.4% full-pool diagnostic. This correspondence is descriptive: the retained summaries do not identify their sample-level overlap. The 8B entry is a single final-checkpoint diagnostic, not a before/after trajectory.

Let $C = \{a_r \neq a^\star\}$. On this subset, $a_p = a^\star$, $a_p = a_r$, and $a_p \notin \{a^\star, a_r\}$ distinguish correct answers, agreement with the incorrect anchor, and other answers; parse failures require separate accounting. Thus not following an incorrect anchor does not generally imply answering correctly. The retained summary reports the two conditional columns separately, without an item-level event partition and denominator reconciliation. Lower wrong-anchor following at 8B is an association across settings, not evidence that model capacity alone prevents error transmission.

The reported wrong-anchor follow rates with and without KL are 32% and 39%† (Table 8). If the two models generate different anchors, they can also induce different conditioning sets C. The available summaries do not establish a common diagnostic cohort and anchor construction, so this difference is treated descriptively rather than as an isolated KL effect on error transmission. The four benchmark-score ablations provide the more direct evidence for regularization.

F.3 Reward and Role-Dynamics Diagnostics

The full-pool and fixed-subset role scores are listed separately in Table 7. Both roles use shared parameters, so a perception-side update can change later reasoning anchors. The observed improvement of both roles is consistent with a useful evolving signal; it does not establish that every update is beneficial or that refreshing outperforms frozen or EMA targets.

For a rollout group g, define

$$Z_g = \mathbb{I}[\mathrm{std}(R_{g,1}, \ldots, R_{g,K}) = 0], \qquad \widehat{z} = \frac{1}{G} \sum_{g=1}^{G} Z_g. \tag{6}$$

The aggregate diagnostics report complementary rates of 22.4% and 77.6%†. At the recorded level of aggregation, nonzero rewards and nonzero within-group reward variance are not distinguished. The percentages are therefore presented as descriptive training diagnostics, not estimates of $\widehat{z}$, $1 - \widehat{z}$, or the fraction of informative GRPO updates; those interpretations would require group-level reward logs beyond the retained summary.

The summary does not identify whether the denominator includes all query-visits or only parser-valid/filtered groups, or how groups are weighted. A group with identical positive rewards has zero centered advantage. Conversely, nonzero variance permits a relative update but does not certify the target's correctness.

The reported mean-reward summary gives endpoints of approximately 0.2 and 0.7. Table 7 lists all four fixed-1k observations in checkpoint order, rather than by exact update index. Figure 3b instead shows the distinct full-6k-pool endpoints. Both sets belong to the RL pool, not a held-out validation

set. The reported experiments use the fixed rollout count; rollout-count sensitivity is outside their scope.

F.4 Repeated-Run Summary and Interpretation

Table 9 contains the three-run Full-PASA mean and SD, distinct from the individual-run main results. The retained record contains these aggregate statistics rather than per-run scores, seed identities, or the SD convention; it contains no three-run OCRBench summary. The variability characterizes PASA across the reported runs, rather than testing significance against single-run baselines.

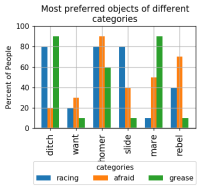
Training-seed variation and evaluation-sample uncertainty answer different questions. A paired bootstrap over benchmark questions would quantify the latter, not the former. Accordingly, the small observed differences between answer-masked and default PASA are interpreted descriptively rather than as an equivalence test. The analysis is confined to the reported summaries, without additional confidence intervals or pairwise significance claims.

F.5 Qualitative Limits of Structured Traces

Figure 5 illustrates a correct answer accompanied by incomplete or inaccurate intermediate evidence. The perception trace lists object names without the numerical values needed for the question, and its description of the bar grouping is inaccurate. The reasoning trace also transcribes the category *afraid* as *affair*. Both outputs nevertheless reach the correct answer, 0. Thus a structured format and answer correctness do not certify faithful visual extraction.

Question: How many objects are preferred by more than 90 percent of people in at least one category?

Answer: 0



<caption>\n### Image 1: Data Plot / Graph**\n\n**Scene Summary:** The image is a bar chart titled \"Most preferred objects of different categories,\" showing the percentage of people's preferences for various objects across three categories: \"racing,\" \"afraid,\" and \"grease.\"\n\n**Explicit Component Breakdown:**\n* **ditch (`ditch`):** A bar representing the percentage of people who prefer \"ditch\" in the \"racing\" category.\n* **want (`want`):** A bar representing the percentage of people who prefer \"want\" in the \"racing\" category.\n* **homer (`homer`):** A bar representing the percentage of people who prefer \"homer\" in the \"racing\" category.\n* **slide (`slide`):** A bar representing the percentage of people who prefer \"slide\" in the \"racing\" category.\n* **mare (`mare`):** A bar representing the percentage of people who prefer \"mare\" in the \"racing\" category.\n* **rebel (`rebel`):** A bar representing the percentage of people who prefer \"rebel\" in the \"racing\" category.\n\n**Interactions and Relationships:** The bars are grouped by category, with each category having three bars representing different objects.\n\n**Implicit and Inferred Properties:**\n* **ditch (`ditch`):** Prefers the \"ditch\" object.\n* **want (`want`):** Prefers the \"want\" object.\n* **homer (`homer`):** Prefers the \"homer\" object.\n* **slide (`slide`):** Prefers the \"slide\" object.\n* **mare (`mare`):** Prefers the \"mare\" object.\n* **rebel (`rebel`):** Prefers the \"rebel\" object.\n\n**Identified Ambiguities:** None.\n</caption>\n\n**Question Analysis:** The question asks how many objects are preferred by more than 90 percent of people in at least one category.\n\n**Solution Process:** \n1. Examine each object's preference across the categories.\n2. Count the objects that have a preference of more than 90 percent in at least one category.\n\n**Answer:** None of the objects are preferred by more than 90 percent of people in any category.\n\n**Final Answer:** \\boxed{0}

<think>\nOkay, let's see. The question is asking how many objects are preferred by more than 90 percent of people in at least one category. The data is presented in a bar graph with three categories: racing, affair, and grease. Each object (ditch, want, homer, slide, mare, rebel) has bars for each category.\n\nFirst, I need to check each object's bars to see if any of them exceed 90% in any category. Let's go one by one.\n\nStarting with \"ditch\": The racing bar is 80%, affair is 20%, grease is 90%. So, the grease is exactly 90%, which is not more than 90, so that doesn't count.\n\nNext, \"want\": Racing is 20%, affair is 30%, grease is 10%. All are under 90%, so no.\n\n\"Homer\": Racing is 80%, affair is 90%, grease is 60%. The affair is exactly 90%, which isn't over, so no.\n\n\"Slide\": Racing is 80%, affair is 40%, grease is 10%. All under 90% except grease, which is 10% here. So no.\n\n\"Mare\": Racing is 10%, affair is 50%, grease is 90%. Grease is exactly 90%, so no.\n\n\"Rebel\": Racing is 40%, affair is 70%, grease is 10%. All under 90% here.\n\nWait, so none of the objects have any category over 90%? But the question says \"at least one category\". Let me double-check each one.\n\nDitch: grease is 90% exactly. So not over.\n\nHomer: affair is 90%, which is not over.\n\nMare: grease is 90%, same.\n\nSo none of the objects have any category exceeding 90%. Therefore, the answer should be 0.\n\nBut wait, maybe I misread the graph. Let me check again. Maybe I made a mistake with the numbers.\n\nLooking at \"homer\": the bars are racing 80, affair 90, grease 60. So affair is exactly 90. So not over.\n\n\"Mare\": grease is 90. So no.\n\n\"Ditch\": grease is 90.\n\nSo all of them are exactly 90 or under. Therefore, no objects have more than 90% in any category. So the answer is 0.\n</think>\n\n \\boxed{0}

Figure 5: **Correct answers can coexist with flawed visual traces.** The perception output omits relevant numerical evidence and misdescribes the grouping; the reasoning output misreads a category label. Outputs, including their formatting artifacts, are shown unchanged. This is a qualitative limitation, not an estimate of error frequency.

This single example does not establish a population-level failure rate or a causal relationship between the trace and the answer. Quantitative evidence-quality comparisons instead use the answer-blind audit in Table 6, within the interpretation scope described in Appendix E.

F.6 Heterogeneous Contextual Comparison

Table 15 provides context across systems that differ in backbone, scale, training data, prompts, and post-training recipes. It is not a controlled ranking and does not establish universal superiority. In particular, its MM-UPT row is distinct from the primary-pair reruns in Table 2.

Table 15: **Non-controlled contextual scores.** I denotes Instruct. Architectures and citations are listed separately in Table 16; the mixed settings are not a controlled ranking.

Model	MMMU	M.Vision	M.Vista	MMStar	OCR	Hall.
General MLLMs (I = Instruct)						
Qwen3-VL-I-2B	53.4	31.6	61.3	58.3	85.8	51.4
InternVL2.5-I-2B	51.3	14.7	51.3	53.7	80.4	42.6
Qwen2.5-VL-I-3B	51.2	21.2	61.2	55.9	79.7	46.3
Gemma3-I-4B	48.8	–	50.0	–	–	–
LLaVA-OV1.5-I-4B	52.7	24.2	67.9	64.9	80.0	–
Phi4-MM-5.6B	55.1	–	62.4	–	84.4	–
Qwen2.5-VL-I-7B	55.0	25.4	67.8	63.6	86.4	–
InternVL2.5-I-8B	56.0	19.7	64.4	62.8	82.2	–
Supervised-RL models						
InternVL3.5-2B	59.0	42.8	71.8	62.7	83.6	48.6
Qwen3-VL-2B	61.4	45.9	73.6	68.1	79.2	54.9
Visionary-R1-3B	30.6	24.7	69.4	66.5	–	–
Vision-SR1-3B	49.6	38.5	–	–	–	–
Source + gold GRPO	56.9	40.8	65.0	56.4	82.7	47.7
Unsupervised/post-training						
MM-UPT	56.0	35.3	66.7	55.9	81.7	47.4
PASA	59.2	41.6	69.5	56.6	83.2	48.1

Table 16: **Architectures behind the contextual comparison.** Model names and language backbones corresponding to Table 15.

Model	Base LLM
Qwen3-VL-Instruct-2B (Bai et al., 2025a)	Qwen3-1.7B
InternVL2.5-Instruct-2B (Chen et al., 2024b)	InternLM2.5-1.8B-Chat
Qwen2.5-VL-Instruct-3B (Bai et al., 2025b)	Qwen2.5-3B
Gemma-3-Instruct-4B (Team et al., 2025)	Gemma-4B
LLaVA-OV1.5-Instruct-4B (An et al., 2025)	Qwen3-4B
Phi-4-Multimodal-5.6B (Abouelenin et al., 2025)	Phi-4-Mini-3.8B
Qwen2.5-VL-Instruct-7B (Bai et al., 2025b)	Qwen2.5-7B
InternVL2.5-Instruct-8B (Chen et al., 2024b)	InternLM2.5-7B-Chat
InternVL3.5-2B (Wang et al., 2025)	Qwen3-1.7B
Qwen3-VL-2B (Bai et al., 2025a)	Qwen3-1.7B
Visionary-R1-3B (Xia et al., 2025)	Qwen2.5-3B
Vision-SR1-3B (Li et al., 2025)	Qwen2.5-3B
Source + gold GRPO	Qwen3-1.7B
MM-UPT (Wei et al., 2025a)	Qwen3-1.7B
PASA	Qwen3-1.7B

Ethics Statement

Improved visually grounded reasoning may benefit educational, document-analysis, and accessibility applications. The same capabilities can also be misused for surveillance, automated decisions, or persuasive explanations unsupported by images. PASA does not provide safety alignment, fairness, privacy protection, or deployment certification. Data/model licenses, potential personal information, and external trace-generation and judging terms remain subject to author-side compliance review. High-stakes deployment requires domain-specific assessment and human oversight; the reported benchmarks are not a substitute.

Reproducibility Statement

The method, trace schema, judge protocol, benchmark tables, and reported merge settings are provided in Sections 3–4 and the appendices. Appendices C, E, and F.4 distinguish the described protocols from the retained execution records. Exact run replication additionally requires model/data revisions, resolved configurations, per-run statistics, and evaluation metadata, which are not fully represented in the retained material.