PHYSALIGN: A BENCHMARK FOR EVIDENCE-GROUNDED ROLE ALIGNMENT IN MULTIMODAL PHYSICS REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

A key challenge in physics diagram understanding is correctly associating visual information with the physical entities, relations, and conditions it describes. Even when a value, symbol, or other local element is accurately recognized, assigning it to the wrong entity or scope can distort the underlying physical premise and lead to incorrect reasoning. To systematically study this challenge, we introduce **PhysAlign**, a benchmark designed to assess whether multimodal models correctly associate information recognized from physics diagrams with its intended physical role. By disentangling visual recognition from physical-role assignment through localized probes and controlled variants, PhysAlign isolates correspondence errors from recognition failures. It contains 3,341 human-validated probes spanning 986 physics problems, enabling systematic evaluation of visual recognition and physical-role correspondence at scale. We further introduce five complementary evaluation metrics, including CAcc, GAcc, and JAcc, which provide a comprehensive assessment of models' ability to recognize diagram content, establish correct physical correspondences, and solve the underlying physics problem. Across our evaluated multimodal models, PhysAlign reveals a consistent gap between local visual recognition and physical-role grounding. Even when the queried content is correctly recognized, the conditional correspondence error rate remains 13.8% for GPT-6-Astra and rises to about 50.6% for InternVL3.5-8B. These findings indicate that strong perception alone does not ensure reliable physical interpretation, exposing a distinct grounding bottleneck that is largely hidden by answer-level accuracy and highlighting the need for future models to better align recognized visual evidence with its physical meaning.

1 INTRODUCTION

Physics reasoning tests whether multimodal large language models (MLLMs) can interpret a physical situation from text and diagrams and apply the appropriate physical laws. A diagram specifies relations among objects and associates quantities with particular entities, while the accompanying text provides further conditions and constraints. These correspondences determine the physical premises on which a solution rests. Evaluating physics reasoning therefore requires examining whether a model recovers the stated problem setup as well as whether it derives the correct answer.

Recent benchmarks examine both the use of visual evidence and intermediate steps in physics reasoning. SeePhys (Xiang et al., 2026a) distinguishes problems that require visual information from those that can be solved from text alone. PhysicsArena (Dai et al., 2025) evaluates variable identification, physical process formulation, and solution derivation, while SeePhys Pro (Xiang et al., 2026b) studies visual variable grounding through semantically aligned input variants that transfer information from text to images. However, diagnosing errors in the recovered problem setup requires a more localized distinction between reading content and assigning it to the correct physical role. Final-answer accuracy cannot determine whether an error originates in recognition, role assignment, or subsequent reasoning. Reconstructing a complete physical graph can also entangle an incorrect correspondence with unrelated entity omissions, naming differences, and structured-output errors. We focus on measuring recognition and role assignment separately at the same evidence location.

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.

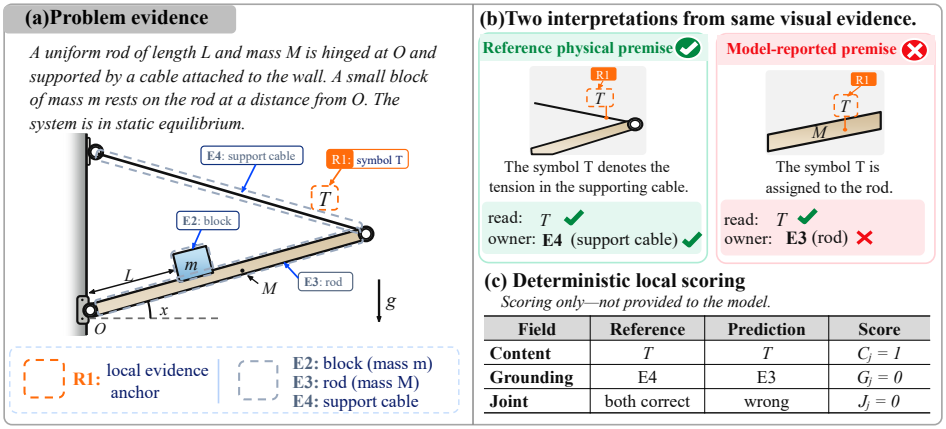

Figure 1: **Correct recognition does not guarantee physical-role grounding.** (a) The problem evidence specifies a rod supported by a cable, with `R1` marking the symbol T. (b) The reference associates T with the supporting cable (`E4`), whereas the illustrated response assigns it to the rod (`E3`). (c) Recognition is correct ($C_j = 1$), but grounding and joint correctness fail ($G_j = 0$, $J_j = 0$). Reference targets and scores are used only for evaluation and are not provided to the model.

Correct recognition of a symbol does not necessarily imply correct grounding to the corresponding physical entity. As illustrated in Figure 1, a rod carrying a block is held in static equilibrium by a hinge at O and a supporting cable. The symbol T at evidence anchor `R1` denotes the tension in the supporting cable (`E4`). The illustrated response recognizes T correctly but assigns it to the rod (`E3`). Here, the queried correspondence refers to the cable associated with the tension symbol, rather than to the body on which the tension force acts. We call the assignment of local content to the physical entity or role supported by the problem evidence *physical-role grounding*. This distinction motivates testing whether role-assignment errors persist when the evidence location is given and even when the correct local reading is supplied.

We introduce **PhysAlign**, a benchmark for diagnosing physical-role grounding through localized probes with explicit evidence anchors. PhysAlign contains 3,341 probes constructed from 986 physics problems across six source datasets. We convert human-reviewed observational annotations into tests of entity reference and physical-quantity ownership. Each probe retains the original problem statement, diagram, and physical conditions, and uses an image region or text span to identify the queried evidence. For each probe, the model identifies the physical role associated with the queried content and, when an independent reading target is available, also reports the corresponding local content. This design enables recognition and role assignment to be evaluated against their respective evidence without requiring full diagram-graph reconstruction.

PhysAlign combines separate recognition and grounding scores with a paired reading control. For probes with independently defined reading and role targets, we use deterministic rules to score recognition and grounding separately; joint correctness requires both targets to be correct. These scores distinguish correctly recognized content from a correctly reported physical premise. A subset of 553 probes also supports paired input variants: the original problem, evidence anchor, question, candidates, and scoring rules remain fixed, while the augmented condition supplies predefined correct local readings. We select these readings using model-independent rules that prohibit disclosure of the queried role correspondence. The comparison measures whether supplying correct content improves grounding. This decomposition concerns observable responses and does not assume separate recognition and grounding modules within a model.

Results from six MLLMs of different scales and model families show that grounding errors persist even when local content is recognized correctly. On the joint evaluation set, the lowest grounding error rate conditional on correct recognition is approximately 13.8%. Paired grounding point estimates increase by 1.3–8.1 percentage points on the observed model-specific subsets, but item-level analysis of four open-weight models reveals both repairs and harms. The highest observed +GT accuracy is 84.9%. These findings support evaluating reported physical premises against their source evidence alongside original-problem solving, without treating supplied readings as perfect perception.

Our contributions are as follows: **(1)** We introduce **PhysAlign**, a benchmark for local physical-role grounding that links tests of entity reference and physical-quantity ownership to explicit source

evidence and human-reviewed role targets. **(2)** We develop deterministic recognition, grounding, and joint scores, together with a paired reading control that tests the effect of supplying correct local content without disclosing the target role correspondence. **(3)** Across model families and scales, we identify correctly recognized content with incorrect physical-role assignments. This discrepancy persists on probes where a frozen proximity rule fails, while paired readings reveal both repairs and harms. These findings expose response-level failures that final-answer accuracy alone cannot localize.

2 Related Work

Multimodal scientific and physics reasoning. Scientific reasoning benchmarks span science question answering in ScienceQA (Lu et al., 2022), multidisciplinary reasoning in MMMU (Yue et al., 2024), visual mathematics in MathVista (Lu et al., 2024), and undergraduate physics in PhysUniBench (Wang et al., 2025a). SeePhys (Xiang et al., 2026a) distinguishes *Vision-Essential* from *Vision-Optional* problems, while SeePhys Pro (Xiang et al., 2026b) studies robustness as task-critical information moves from text to diagrams. PhysicsArena (Dai et al., 2025) further evaluates intermediate stages such as variable identification and process construction. These benchmarks primarily assess final solutions or broad reasoning stages, where perception, entity correspondence, and derivation remain coupled. PhysAlign instead examines the pre-reasoning recovery of the problem setup: whether content read from a local visual region is assigned to the correct physical entity or role.

Controlled visual diagnosis and role grounding. Controlled vision–language evaluations show that recognizing individual elements does not ensure correct relational interpretation. Winoground (Thrush et al., 2022), ARO (Yuksekgonul et al., 2023), VL-CheckList (Zhao et al., 2022), and SugarCrepe (Hsieh et al., 2023) isolate failures in compositional structure, attributes, relations, and order. Related diagnostics for mathematical and scientific reasoning vary the availability or modality of visual evidence: MathVerse (Zhang et al., 2024) changes the visual–textual composition of matched problems, MMStar (Chen et al., 2024) emphasizes vision-essential evaluation and textual-shortcut control, and SeePhys Pro (Xiang et al., 2026b) probes modality transfer and visual variable grounding. PhysAlign instead fixes the original problem and evidence anchors, separately measures *Content Recognition* and *Role Grounding*, and uses paired controls that supply correct local readings without revealing the grounding target. A protocol-level comparison with related benchmarks is provided in Appendix E (Table 18).

3 PhysAlign: An Evidence-Grounded Role Alignment Benchmark

3.1 Task Definition and Scope

Each PhysAlign probe evaluates a local physical correspondence using the original problem text and diagram, an evidence anchor, a local question, and candidates identified by neutral aliases. The anchor specifies an image region or an exact text span. The model selects the corresponding physical role target and, when a separate reading target is defined, reports the content at the anchor.

In Figure 1, anchor `R1` marks the tension symbol T, and the queried correspondence is the supporting cable (`E4`). A response with `read`: T and `owner`: `E3` correctly recognizes the symbol but incorrectly assigns it to the rod. For this query, the target is the cable associated with the tension, rather than the body on which the force acts.

The benchmark uses three probe types: `T02-single` resolves an anchored mention to an entity; `T03-image` transcribes an anchored quantity label and identifies its owner; and `T03-text` identifies the owner of an anchored textual quantity. Each query preserves the relation specified by its source annotation, and a quantity's owner may be an object, force, or field. Joint evaluation requires separately defined targets for content recognition and role grounding; probes without an independent reading target receive grounding scores only. This decomposition concerns observable responses and makes no assumption about separate recognition and grounding modules within a model.

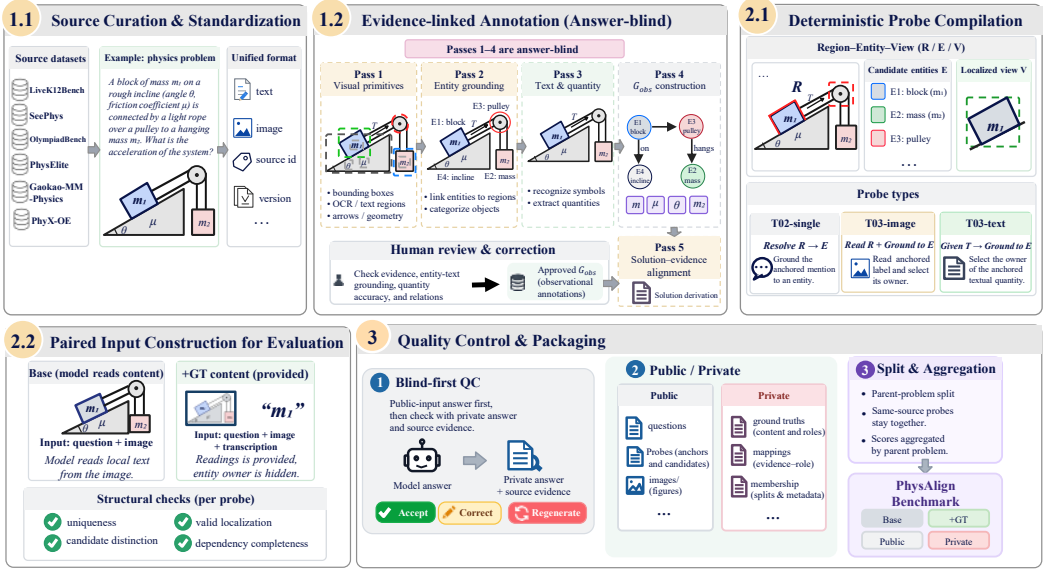

Figure 2: **Data construction and quality assurance in PhysAlign.** Standardized source problems undergo answer-blind annotation (Passes 1–4), human review, and probe compilation. Pass 5 aligns reference solutions with evidence for analysis only and does not supply probe targets. Paired base/+GT inputs share the original problem, evidence anchors, questions, and candidates; +GT supplies correct local readings without disclosing the queried role target. Structural checks and instance audits precede packaging, which separates public inputs from private targets and scoring mappings.

3.2 Evidence-Linked Annotations and Probe Construction

We construct probes from reviewed annotations linked to the original problem evidence. Figure 2 summarizes the pipeline, and Appendix A provides provenance records and admission criteria.

We use GPT-5.6 Sol (OpenAI, 2026a) with high reasoning effort to draft structured annotations, which are then reviewed and corrected against the original text and diagrams. Passes 1–4 record visible primitives, entity references, explicit quantities and conditions, and their supporting evidence without access to reference answers or solutions. Only these reviewed observational records supply probe targets. Pass 5 separately aligns reference solutions with evidence for analysis.

Fixed templates convert supported annotation fields into local questions. Candidates use neutral aliases linked to image regions or text spans. Alias assignments and candidate order are randomized once and held fixed across models. Single-target probes require at least two valid, type-compatible candidates; set-valued templates retain all correct targets. Candidate descriptions exclude answer-bearing role associations. Reference readings, role targets, alias mappings, and source-annotation links are stored separately from public inputs. Eligible probes also receive the paired +GT input defined in Section 4.1.

Quality assurance comprises source-annotation review, model-assisted probe screening, and human verification of exported probes. Checks assess evidence localization, question semantics, candidate completeness and distinguishability, target support, and information leakage. During human verification, two coauthors independently answer the public probe before consulting private targets and source annotations; a third reviewer adjudicates disagreements. Ambiguous or unsupported records are withheld pending resolution. Appendix A.3 details construction-stage verification; Appendix A.5 reports the post-construction audit.

3.3 Dataset Scale and Coverage

PhysAlign contains **986 parent problems** from six sources and **3,341 unique localized probes**; **553 probes** from 385 parents also support paired +GT inputs, without adding unique probes. Table 1 reports source coverage and the parent-disjoint release partition. Evaluation pools are drawn from the full release (development and test combined); metric-specific eligibility and observed coverage are defined in Section 4.2.

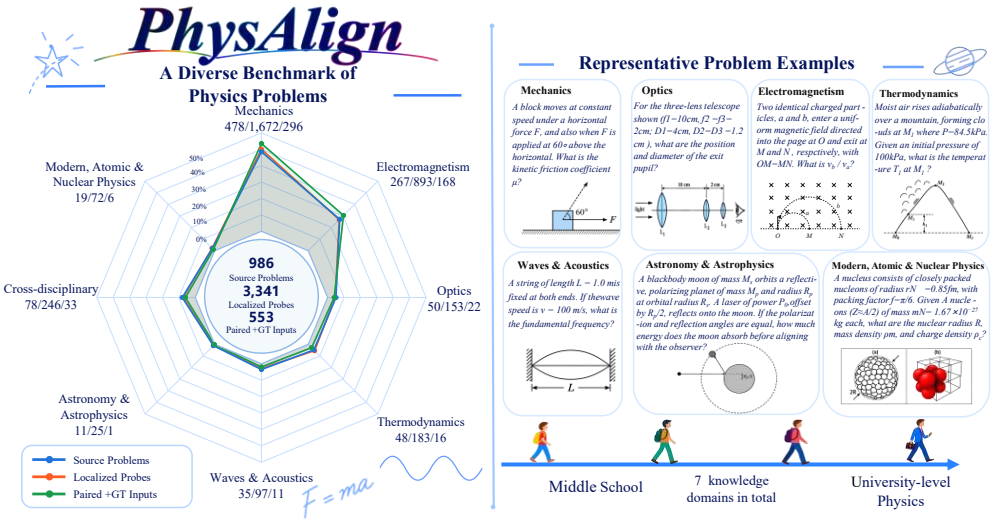

Figure 3: **Domain coverage and representative physics problems in PhysAlign. Left:** Distributions of source problems, localized probes, and paired +GT inputs across seven physics domains and a cross-disciplinary group. Each radar series shows percentages within its corresponding collection; counts beside each group follow the same order. **Right:** Representative question–diagram examples from the seven domains, illustrating an educational span from middle school to university-level physics.

Table 1: Source coverage and release partition of PhysAlign.

(a) Source coverage

Source dataset	Parents	Parents with +GT
SeePhys (Xiang et al., 2026a)	268	119
LiveK12Bench (Wang et al., 2026)	250	88
PhysElite (Xu et al., 2026)	220	79
OlympiadBench[a] (He et al., 2024)	209	68
Gaokao-MM-Physics (Zong & Qiu, 2024)	23	17
PhyX-OE (Shen et al., 2025)	16	14
Full release	**986**	**385**

[a] English physics competition subset.

(b) Development/test partition

Split	Parents	Local probes	Paired +GT
Development	79	208	90
Test	907	3,133	463
Full release	**986**	**3,341**	**553**

Evaluation. Both splits are used jointly. Paired +GT inputs are variants, not additional unique probes.

The source problems span mechanics, electromagnetism, optics, thermal physics, waves, modern physics, and astrophysics (Figure 3). Mechanics and electromagnetism account for most parent problems under the estimated domain assignments. The illustrated problems provide context for the local correspondences evaluated by PhysAlign. Appendix A.6 reports estimated domain composition, probe density, and paired coverage, with parent-level and probe-level statistics reported separately.

4 Evaluation Protocol and Metrics

4.1 Paired Evaluation with Ground-Truth Local Content

For each eligible probe, the *Base* input contains the original problem, local question, evidence anchors, and candidates. The matched *+GT* input additionally supplies verified ground-truth (GT) readings of the specified local content while withholding the queried physical role. Both conditions use the same question, candidates and aliases, output schema, inference settings, and scoring rules, and are run in independent sessions without shared history, predictions, or feedback. Supplied readings are inputs and are therefore excluded from recognition scoring.

Model-independent rules determine which readings may be supplied. In Figure 1, +GT may state that R1 reads T, but not that T denotes the tension in the supporting cable (E4), since the latter reveals the grounding target. Candidate-side information, when present, is selected symmetrically across candidates. The paired comparison measures how grounding changes when verified local content is supplied. It does not provide a complete physical description or isolate the causal effect of correcting OCR errors.

4.2 Deterministic Scoring and Balanced Aggregation

Let $\mathcal{G}$ contain all 3,341 grounding probes from 986 parents. The joint set $\mathcal{L} \subseteq \mathcal{G}$ contains 412 probes from 311 parents with independent reading targets; the paired-eligible release set $\mathcal{P} \subseteq \mathcal{G}$ contains 553 probes from 385 parents with valid +GT inputs. Recognition and joint metrics use $\mathcal{L}$, and overall grounding uses $\mathcal{G}$. Base/+GT comparisons use model m's `T03-image` probes with both response records available, $\mathcal{P}_m \subseteq \mathcal{L} \cap \mathcal{P}$ (Appendix B.5).

For predicted content $\hat{c}_j$ and candidate identifier $\hat{u}_j$, let c_j and r_j denote the reference content and role target. Given the fixed content normalizer N_j and candidate mapping ϕ_j,

$$C_j = \mathbf{1}[N_j(\hat{c}_j) = N_j(c_j)], \quad G_j = \mathbf{1}[\phi_j(\hat{u}_j) = r_j], \quad J_j = C_j G_j. \tag{1}$$

C_j and J_j are defined on $\mathcal{L}$, whereas G_j applies to $\mathcal{G}$. Content must match the specified anchor rather than an unordered inventory. Grounding is scored independently, while J_j requires correct recognition and grounding for the same local premise. Base and +GT use the same grounding rule. Normalization, multi-field and set matching, and format-error handling are fixed before evaluation; these local scores use no open-ended language-model judge.

For an evaluation set $\mathcal{S}$, $\mathrm{Agg}_{\mathcal{S}}$ first averages probes within each parent–task-type group, then parents within each task type, and finally task types using prespecified weights (see Appendix B.3 for the formal definition). We report

$$CAcc = \mathrm{Agg}_{\mathcal{L}}(C), \quad JAcc = \mathrm{Agg}_{\mathcal{L}}(J), \quad GAcc^{v}_{\mathcal{S}} = \mathrm{Agg}_{\mathcal{S}}(G^{v}), \quad v \in \{\mathrm{base}, +\mathrm{GT}\}. \tag{2}$$

This aggregation prevents annotation-rich parents from dominating a task type. Overall grounding is $GAcc^{\mathrm{base}}_{\mathcal{G}}$. Paired scores use identical Base/+GT support and weights within each $\mathcal{P}_m$, with $\Delta G_{\mathcal{P}_m} = GAcc^{+\mathrm{GT}}_{\mathcal{P}_m} - GAcc^{\mathrm{base}}_{\mathcal{P}_m}$; +GT performance is not assumed to be an upper bound.

On $\mathcal{L}$, $C^{+}G^{-}$ denotes the event $C_j = 1, G_j = 0$, i.e., correct content recognition with incorrect grounding; the other sign combinations are defined analogously. Under the same aggregation, its weighted share is $CAcc - JAcc$, not $CAcc - GAcc$, and $JAcc$ is not the product of the marginal accuracies. We also report $GAcc^{\mathrm{base}}_{\mathcal{L}}$ for same-set comparisons.

5 Experiments and Diagnostic Analysis

5.1 Experimental Setup and Main Results

We evaluate Qwen3.5-4B, -9B, and -27B (Team, 2026), InternVL3.5-8B (Wang et al., 2025b), GPT-6 Astra (OpenAI, 2026b), and Gemini-3.8-flash (Google DeepMind, 2026) using pools drawn from both release partitions, not test alone. Table 2 reports local metrics on $\mathcal{L}$, $\mathcal{G}$, and $\mathcal{P}_m$, and independent solving on all 986 parents. $SolveAcc$ averages normalized problem credit: scripted binary credit for choice and fill-in items, and expert-assigned partial credit for short-answer and calculation items (Appendix B.4). Local metrics use balanced weights; solving weights parents equally. Repeated Base scores reflect shared probes and weights; API deviations reflect response coverage (Appendix B.5). Inference settings are in Appendix C.1.

Two patterns emerge from the aggregate results. Within the Qwen3.5 family, scaling from 4B to 27B coincides with substantial gains in both content recognition and joint performance: $CAcc$ increases from 53.97% to 84.14%, while $JAcc$ rises from 32.09% to 68.12%. Yet stronger content recognition does not eliminate the gap between reading local content and grounding its physical role. On the same joint set $\mathcal{L}$, GPT-6 Astra (OpenAI, 2026b) obtains 89.44% $CAcc$, 83.31% $GAcc^{\mathrm{base}}_{\mathcal{L}}$, and 77.13% $JAcc$; the corresponding results for Gemini-3.8-flash are 85.20%, 79.33%, and 70.59%. Thus, gains in scale and aggregate accuracy do not by themselves ensure consistent recovery of local content and its physical role. Marginal accuracies also cannot reveal whether a grounding error follows a recognition error or occurs after the content has already been read correctly. We therefore decompose both outcomes on the same local probes.

5.2 Recognition–Grounding Decomposition and Local-Reading Controls

Figure 4(a) decomposes the joint set $\mathcal{L}$ into four outcomes according to whether content recognition and grounding are correct. The central result is that correctly reading the content at a specified location

Table 2: Recognition, grounding, paired local-reading controls, and independent original-problem solving.

	Joint ($\mathcal{L}$)			All ($\mathcal{G}$)	Reading control ($\mathcal{P}_m$)			Solving
Model	$CAcc \uparrow$	$GAcc_{\mathcal{L}}^{\text{base}} \uparrow$	$JAcc \uparrow$	$GAcc \uparrow$	$GAcc_{\mathcal{P}_m}^{\text{base}} \uparrow$	$GAcc_{\mathcal{P}_m}^{+\text{GT}} \uparrow$	$\Delta G_{\mathcal{P}_m}$	$SolveAcc \uparrow$
Qwen3.5-4B	53.97	46.51	32.09	40.21	46.51	54.61	**+8.10**	9.94
Qwen3.5-9B	65.81	54.55	37.55	40.52	54.55	56.53	+1.98	12.12
Qwen3.5-27B	84.14	80.29	68.12	84.11	80.29	82.00	+1.71	18.39
InternVL3.5-8B	2.12	33.36	1.05	23.52	33.36	34.81	+1.45	7.47
GPT-6 Astra	**89.44**	**83.31**	**77.13**	81.94	**83.59**	**84.90**	+1.32	**71.72**
Gemini 3.8 Flash	85.20	79.33	70.59	**90.71**	78.67	80.86	+2.20	21.31

Scores are percentages; $SolveAcc$ is mean normalized credit, not binary accuracy (Appendix B.4). Parent/probe counts: $\mathcal{L}$, 311/412; $\mathcal{G}$, 986/3,341; $\mathcal{P}_m$, 311/412 (open weights), 304/397 (APIs); solving, 986 parents. $GAcc = GAcc_{\mathcal{G}}^{\text{base}}$; $\Delta G_{\mathcal{P}_m}$ is the paired change in percentage points, rounded independently. Coverage details: Appendix B.5. **Blue bold**/underlined red: column maxima/minima; green: positive changes (largest in **bold**), not significance.

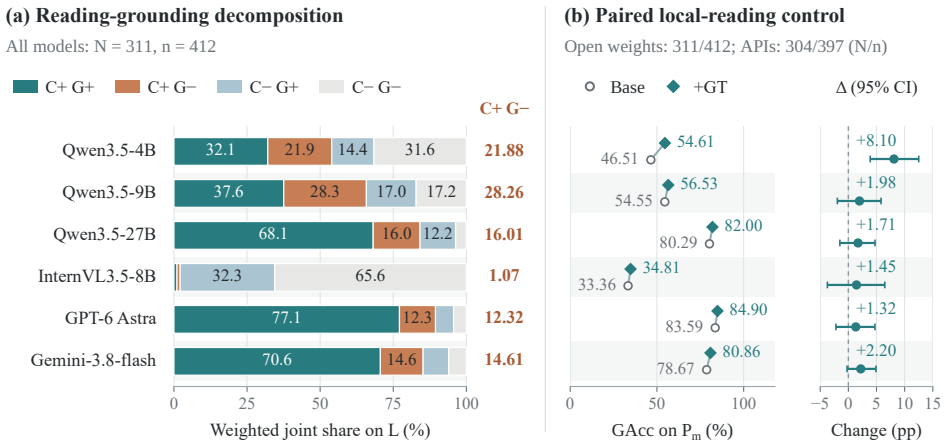

Figure 4: Recognition–grounding decomposition and paired controls. (a) Weighted outcome shares on $\mathcal{L}$ ($N = 311$, $n = 412$). (b) Base/+GT grounding and paired changes with 95% confidence intervals on $\mathcal{P}_m$; model-specific coverage is shown above the panel (Appendix B.5). N/n count parents/probes. +GT supplies local readings, not perfect perception.

does not guarantee correct grounding of its physical role. Excluding InternVL3.5-8B (Wang et al., 2025b), whose local recognition is nearly absent, the weighted share of content-correct but grounding-wrong events (C^+G^-) ranges from 12.32% to 28.26% across the other five models. The shares are 21.88%, 28.26%, 16.01%, 12.32%, and 14.61% for Qwen3.5-4B (Team, 2026), Qwen3.5-9B (Team, 2026), Qwen3.5-27B (Team, 2026), GPT-6 Astra (OpenAI, 2026b), and Gemini-3.8-flash (Google DeepMind, 2026), respectively. Its recurrence across model families and scales shows that this is not an isolated failure: models can read the specified text or value correctly and still assign it to the wrong physical entity or role.

The C^+G^- quantity is a weighted event share over the full joint set, not a conditional grounding error rate given correct recognition. Its magnitude therefore also depends on how often a model reads the content correctly. InternVL3.5-8B (Wang et al., 2025b), for example, has only 1.07% C^+G^- but also only 2.12% $CAcc$; the small event share does not imply more reliable role grounding. The four-way decomposition diagnoses failure modes hidden by marginal accuracy rather than providing another model ranking.

Figure 4(b) reports positive point estimates for Base-to-+GT changes of 1.32–8.10 percentage points on model-specific observed sets $\mathcal{P}_m$. The transition analysis for the four open-weight models (Appendix D.1, Table 12) reveals both repairs and harms. For Qwen3.5-4B, +GT repairs 13.35% and harms 5.25% of the weighted probe mass, yielding a net gain of 8.10 points (95% CI [3.88, 12.50]). The other three net-change intervals include zero. Net gains therefore do not imply uniform item-level benefits.

Even under +GT, Qwen3.5-27B, GPT-6 Astra, and Gemini-3.8-flash retain weighted grounding error rates of 18.00%, 15.10%, and 19.14%, respectively, on their observed $\mathcal{P}_m$. Supplied readings therefore neither guarantee role assignment nor define a grounding upper bound. Since +GT withholds

Table 3: Grounding validity and robustness: rule-based baselines, image removal, and candidate reordering.

Rule-based baselines ($GAcc$ ↑)
Random candidates: 18.69% Nearest region: 76.82%

Information controls and interface robustness

	Reference	No image		Reordered
Model	$GAcc$ ↑	$GAcc$ ↑	ΔG (pp)	$GAcc$ ↑
Qwen3.5-4B (Team, 2026)	40.21	38.78	−1.43	41.93
Qwen3.5-9B (Team, 2026)	40.52	43.53	+3.01	43.02
Qwen3.5-27B (Team, 2026)	84.11	63.72	**−20.39**	84.48
InternVL3.5-8B (Wang et al., 2025b)	23.52	49.87	**+26.35**	7.24
GPT-6 Astra (OpenAI, 2026b)	81.94	75.57	−6.37	84.27
Gemini 3.8 Flash (Google DeepMind, 2026)	**90.71**	**76.90**	−13.81	**90.82**

Accuracies are percentages (↑). Reference is each model's $GAcc$ (all) in Table 2; Reordered permutes candidates. $\Delta G = GAcc_{\text{no image}} - GAcc_{\text{base}}$ is in percentage points, rounded independently. Random is the evaluation-weighted expectation of $1/K_j$ on single-target probes; rule scores here use their original pools. Same-support comparisons appear in Appendix D.2. **Blue bold**/underlined red: model-accuracy maxima/minima; green/red shading: positive/negative changes (extremes in **bold**), not significance or better/worse rankings.

the correct owner or role and leaves other visual and spatial uncertainties unresolved, the paired changes do not isolate a pure perception effect, and residual errors cannot be attributed exclusively to physical reasoning.

5.3 Validity: Visual Evidence, Geometry, and Interface Robustness

Table 3 evaluates three alternative explanations for the grounding results: dependence on visual input, simple geometric regularities, and the candidate interface. Each control addresses a distinct question and is interpreted within its own scope.

Removing the image substantially reduces grounding accuracy for the strongest models. Qwen3.5-27B (Team, 2026), GPT-6 Astra (OpenAI, 2026b), and Gemini-3.8-flash (Google DeepMind, 2026) lose 20.39, 6.37, and 13.81 percentage points, respectively. The consistent declines show that their high reference scores cannot be explained entirely by the text-only interface or candidate priors. Weaker models do not follow the same pattern: Qwen3.5-9B (Team, 2026) and InternVL3.5-8B (Wang et al., 2025b) gain 3.01 and 26.35 points without the image. Visual input therefore interacts strongly with the model and the evaluation interface rather than conferring a uniform benefit. Removing the image can also alter the evidence anchors and the identifiability of candidate targets. This control diagnoses visual-input sensitivity only when the remaining input still defines a valid task; it does not isolate the exact contribution of visual reasoning.

The spatial baseline requires a same-support comparison rather than subtraction from overall $GAcc$. On the fixed applicable subset (665 parents, 1,639 probes), Qwen3.5-27B achieves 86.35% against 76.82% for nearest-region, a 9.53-point margin (95% CI [6.10, 13.06]). We further define *H-hard* as the probes on which the frozen nearest-region rule is incorrect, independently of model predictions (Appendix D.2). On this subset (271 parents, 394 probes), Qwen3.5-27B achieves 73.22% grounding accuracy against 18.76% for same-support random selection. On its joint-reading intersection (69 parents, 74 probes), the model's C^+G^- share remains 29.71% [20.14, 39.84], with a 34.75% [23.77, 46.55] grounding error rate conditional on correct reading. The tested proximity rule thus does not fully explain this model's grounding performance, and reading–grounding disagreement persists where the rule fails. This does not exclude other shortcuts. Candidate reordering changes aggregate $GAcc$ by at most 2.50 percentage points for five models. InternVL3.5-8B (Wang et al., 2025b) is the exception, falling from 23.52% to 7.24%. Thus, five models exhibit limited aggregate sensitivity to the tested reordering, not established permutation invariance. Similar accuracies can conceal offsetting item-level changes; alias preferences and sensitivity to other orderings remain untested.

Together, these controls distinguish sensitivity to image evidence, performance beyond the tested proximity rule, and robustness to the tested candidate reordering. They motivate reporting geometry-defined hard-subset performance alongside aggregate grounding, without identifying internal strategies or establishing a shortcut-free benchmark. We next examine whether successful original-problem solving certifies accurate local grounding.

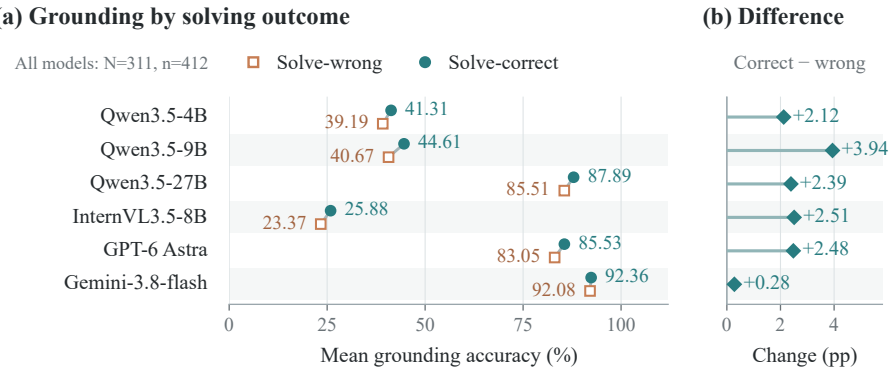

Figure 5: Local grounding by binary final-answer correctness, not by partial-credit $SolveAcc$. (a) Parent-level means for solve-wrong and solve-correct groups ($N = 311$ parents, $n = 412$ probes). (b) Correct-minus-wrong difference in percentage points. Estimates use independent solving and probing runs.

5.4 Association with Original-Problem Solving

We next compare parent-level grounding by binary final-answer correctness, a diagnostic distinct from the partial-credit $SolveAcc$. Figure 5 uses 311 parents and 412 local probes from the full release. Solving and probing use independent contexts; Appendix D.3 details the grouping and aggregation.

All six models have higher grounding point estimates on correctly solved problems, but the differences span only 0.28–3.94 percentage points (Figure 5(b)). This common direction suggests a modest descriptive association, not equivalence between the two tasks. The absolute scores provide the sharper contrast (Figure 5(a)): on correctly solved problems, Qwen3.5-27B (Team, 2026) and GPT-6 Astra (OpenAI, 2026b) still incur local grounding error rates of 12.11% and 14.47%, respectively. Conversely, Gemini-3.8-flash (Google DeepMind, 2026) retains 92.08% grounding accuracy on incorrectly solved problems. Correct answers thus do not certify accurate local reports, and high local grounding accuracy does not certify successful problem solving.

PhysAlign therefore complements solution accuracy by testing reported evidence–role correspondences that final answers do not resolve. This conclusion concerns observable behavior, not the premises used internally: the model-specific grouping and independent contexts neither identify which local evidence entered a solution nor establish that repairing a grounding error would change its answer. Appendix D.3 provides the aggregation details, full estimates, and interpretive boundaries.

6 Conclusion

PhysAlign complements solution accuracy with anchored probes that separate content recognition from physical-role grounding. Our results show that correct recognition does not guarantee grounding, reference readings resolve only some grounding errors, and correct final answers can coexist with incorrect reported premises. These findings motivate models that verify evidence–role correspondences through query-conditioned local checks and are trained with hard negatives that preserve content while swapping entities or roles. Our findings concern explicit reports under fixed anchors and candidates in static diagrams, not complete scene understanding, internal model states, or premises used in a particular solution. Domain coverage is uneven, and whether such targeted training improves grounding and problem solving or transfers to dynamic settings remains to be tested.

AI Use Statement

Generative AI tools assisted with structured annotation drafting, candidate-probe screening, and annotation correction, as well as literature search and language editing. Annotation drafts were reviewed and corrected by the authors against the original text and diagrams. The authors take full responsibility for the dataset, methodology, experimental analyses, conclusions, and final manuscript.

REFERENCES

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.

Song Dai, Yibo Yan, Jiamin Su, Dongfang Zihao, Yubo Gao, Yonghua Hei, Jungang Li, Junyan Zhang, Sicheng Tao, Zhuoran Gao, et al. Physicsarena: The first multimodal physics reasoning benchmark exploring variable, process, and solution dimensions. In *EMNLP (Findings)*, pp. 17290–17316, 2025.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Google DeepMind. Gemini 3.8 Flash: Model card. Google DeepMind Model Card, 2026. URL `https://deepmind.google/models/model-cards/gemini-3-8-flash/`. Published September 2, 2026. Accessed September 16, 2026.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. Ai2d-rst: a multimodal corpus of 1000 primary school science diagrams: T. hiippala et al. *Language Resources and Evaluation*, 55(3):661–688, 2021.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

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*, volume 2024, pp. 23439–23554, 2024.

OpenAI. GPT-5.6 Sol model. OpenAI API Documentation, 2026a. URL `https://developers.openai.com/api/docs/models/gpt-5.6-sol`. Accessed: 2026-09-15.

OpenAI. Gpt-6 astra model. OpenAI API Documentation, 2026b. URL `https://developers.openai.com/api/docs/models/gpt-6-astra`. Accessed: 2026-09-14.

Hui Shen, Taiqiang Wu, Qi Han, Yunta Hsieh, Jizhou Wang, Yuyue Zhang, Yuxin Cheng, Zijian Hao, Yuansheng Ni, Xin Wang, et al. Phyx: Does your model have the" wits" for physical reasoning? *arXiv preprint arXiv:2505.15929*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Textvqa: Towards vqa requiring reasoning about text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 6, 2019.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, June 2022.

Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Peng Xia, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, et al. Physunibench: an undergraduate-level physics reasoning benchmark for multimodal models. *arXiv e-prints*, pp. arXiv–2506, 2025a.

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*, 2025b.

Xiaohan Wang, Mingze Yin, Yilin Zhao, Gang Liu, and Dian Li. Livek12bench: Have large multimodal models truly conquered high school-level examinations? *arXiv preprint arXiv:2605.26781*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, et al. Seephys: Does seeing help thinking?–benchmarking vision-based physics reasoning. *Advances in Neural Information Processing Systems*, 38, 2026a.

Kun Xiang, Terry Jingchen Zhang, Zirong Liu, Bokai Zhou, Yueling Tang, Junjie Yu, Jiacong Lu, Shangrui Huang, Heng Li, Likui Zhang, et al. Seephys pro: Diagnosing modality transfer and blind-training effects in multimodal rlvr for physics reasoning. *arXiv preprint arXiv:2605.09266*, 2026b.

Ruoran Xu, Wending Gao, Liyunfeng Chen, Aixin Shi, Haoyu Cheng, Zixiang Fang, Yiqiang Zou, and Qiufeng Wang. Physelite: How far are llms from solving olympiad-level physics problems? *arXiv preprint arXiv:2608.25097*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023. URL https://arxiv.org/abs/2210.01936.

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.

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16593–16615, 2025.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

Yi Zong and Xipeng Qiu. Gaokao-mm: A chinese human-level benchmark for multimodal models evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8817–8825, 2024.

A Dataset Construction, Annotation, and Quality Assurance

This appendix supplements the construction protocol in Section 3.2 and the dataset overview in Section 3.3. Source coverage and split statistics are reported in Table 1 in the main text; here we provide provenance records, annotation details, review procedures, release accounting, and a post-construction quality audit. Unless an individual split or candidate pool is specified, counts refer to the full retained release, combining development and test. Evaluation pools are drawn from this release; observed paired coverage is distinguished from release eligibility in Appendix B.5. Split-specific counts describe the release inventory, not separate performance evaluations. Section A.5 reports an audit of the retained inventory and the disposition of flagged probes in a subsequent release candidate.

A.1 Source Problems and Provenance

Source records. Table 1 lists the six source datasets and their retained parent counts, rather than the sizes of the upstream datasets. For OlympiadBench (He et al., 2024), we use its English physics competition subset. Each source record preserves the original image, statement, question, answer options when present, reference answer or solution, and provenance identifiers. PhysAlign augments these records with evidence-linked observational annotations, local probes, candidate mappings, recognition packets, and scoring records. The original physics problems are reused rather than newly authored.

Deduplication and split propagation. Multiple probes can originate from the same parent problem and therefore do not represent independent source problems. The merge procedure deduplicates exactly identical public inputs and propagates split assignments and isolation decisions through original-image hashes and existing recorded groups. The current merge introduced no additional exclusions. All retained probes from a parent remain in the same split, with membership recorded in `private/membership.json`. These checks address exact-input duplication and recorded grouping; they do not establish the absence of semantic near-duplicates or prior model exposure to the original problems.

A.2 Observational Annotation and Probe Compilation

Pass-level outputs. The observational stage in Figure 2 comprises four answer-blind passes. Pass 1 records visible primitives, text, and locations; Pass 2 links them to physical entities; Pass 3 records textual mentions, explicit quantities, constraints, and query targets; and Pass 4 consolidates directly supported observations and evidence links into G_{obs}. Correct answers and reference solutions are excluded from these four passes.

Pass 5 aligns atomic steps of an existing reference solution with the observations they use. It is reserved for analysis: it neither re-solves the problem nor supplies probe targets or solution-derived corrections to G_{obs}.

Exported fields and mappings. The interfaces in Section 3.1 export `referent` for `T02-single`, `read` and `owner` for `T03-image`, and `owner` for `T03-text`. Each instance associates an evidence anchor (R), neutral candidate entities (E), and an original or localized view (V). Targets are compiled from reviewed observational fields, while alias-to-entity mappings remain on the scoring side.

Templates preserve the physical meaning of their source fields. In particular, a quantity's `owner_id` identifies the entity to which that quantity belongs; the owner may be an object, force, field, or another physical entity. Identifying the force entity whose magnitude is shown, for example, differs from identifying the object on which that force acts. The compiler does not substitute one relation for the other. Appendix B specifies the input/output interfaces and +GT information rules.

A.3 Quality Control and Admission Rules

Verification and correction. Quality control checks both the source annotations and the exported probes. Source verification compares transcription, localization, entity correspondence, quantity ownership, and evidence support against the original problem. Exported-probe review examines the

complete public question, evidence anchor, candidate set, and scoring target. Structural checks validate references, dependencies, mappings, and localization; semantic correctness, answer uniqueness, and candidate distinguishability require review against the original text and images.

Human verification follows the blind-first procedure in Section 3.2. Two coauthors independently answer the public probe before consulting the private target and source annotations. Direct approval requires agreement between their independent answers and successful quality checks; disagreements proceed to a third reviewer for adjudication. Review also checks candidate completeness, transcription, localization, and information leakage. Corrections require regeneration and rechecking, while shared-source or rule-level issues require examination of related instances. The construction-stage procedures above are distinct from the post-construction audit of retained probes in Section A.5.

Revision of further-review candidates. We used GPT-6 Astra (OpenAI, 2026b) to revise the 1,089 candidates marked as requiring further review at initial screening. The first stage checked probe content against the original text and images and corrected transcription and reading-target errors where applicable. The second stage compared evidence regions, candidate entities, and target relations with the original images to determine the required localization and semantic corrections. The revised probes then underwent item-by-item human verification using the same blind-first procedure. These decisions apply to the revised versions; the original screening states remain part of the initial candidate accounting. Section A.4 reports the final review outcomes and distinguishes the supplementary candidates from the retained benchmark.

Admission and audit interpretation. The recorded construction-stage test-parent admission condition is

$$n_i^{\mathrm{rej}} \leq 2,$$

where n_i^{rej} counts generated candidate probes for parent i recorded as rejected. This condition governs parent eligibility; it neither admits rejected probes nor overrides subsequent risk isolation. Candidates with unresolved screening states or incomplete calls were excluded from the reported benchmark. Subsequent revisions and their review outcomes are accounted for separately.

Table 4(a) records model-screening states, whereas panel (c) reports final human-review outcomes for revised probes. Model approval does not establish agreement between human reviewers, and a further-review state does not itself indicate human disagreement. Neither the screening counts nor final post-revision approvals are used to infer pre-adjudication agreement or a human-performance ceiling.

Input–scoring separation. Model-visible files contain base and +GT inputs together with the images available for inference. Reference targets, source and candidate mappings, parent membership, source snapshots, and review records are stored separately for scoring and audit.

A.4 Candidate Screening and Release Retention

Table 4(a) accounts for 5,594 generated probes from 1,084 candidate parents. Initial model screening yielded 3,370 approvals, 1,128 rejections, 1,089 candidates requiring further review, and seven incomplete or failed screening calls. These four states are mutually exclusive. Risk isolation subsequently excluded 29 model-approved candidates, leaving 3,341 retained probes from 986 parents, as shown in panel (b). The retained probes constitute 59.72% of generated candidates. The other 2,253 candidates were not admitted to the reported benchmark at that stage: 1,128 model rejections, 1,089 further-review cases, seven incomplete calls, and 29 risk-isolated approvals. These initial states do not imply that every excluded candidate was semantically incorrect; subsequent revisions are reported separately.

Among the 986 retained parents, 184 retain their complete generated probe set, whereas 802 retain only an approved subset. Here, *complete* means that all generated candidates for the parent are retained, not that every physical fact in the problem is annotated. Partial retention means that only probes satisfying admission requirements enter the benchmark; candidates rejected or unresolved at screening were not admitted in those versions. The remaining $1{,}084 - 986 = 98$ candidate parents contribute no retained probe. The available accounting does not attribute this parent-level difference to a single exclusion cause.

Table 4: **Candidate-probe screening, release admission, and human re-review.** Panel (a) records the initial model-screening states of 5,594 generated probes; panel (b) reports admission within the model-approved subset. Panel (c) reports final human-review outcomes for revised versions of the 1,089 candidates initially requiring further review.

Recorded status or outcome	Probes
(a) Initial model-screening states	
Approved by model screening	3,370
Rejected by model screening	1,128
Further review required at initial screening	1,089
Incomplete / call error	7
Total generated candidates	**5,594**
(b) Admission within model-approved candidates	
Withheld through risk isolation	29
Retained probes	**3,341**
(c) Human re-review of revised candidates	
Revised probes reviewed by humans	1,089
Approved after review; retained in supplementary pool	1,089
Rejected or unresolved after review	0
Added to the reported benchmark	0

The states in (a) sum to 5,594. The 29 isolated candidates are included in the 3,370 approvals, so (b) sums to $29 + 3{,}341 = 3{,}370$; panels must not be added together. Panel (c) follows revised versions of the 1,089 further-review candidates in (a). Its coverage and outcome rows describe the same probes and must not be summed. These revised probes form a separate supplementary candidate pool and were not used in the reported evaluation. Human outcomes are final post-revision decisions, not first-pass acceptance or inter-reviewer agreement. Incomplete calls concern construction-stage screening.

Human re-review and supplementary candidates. Panel (c) reports the outcomes of the revision procedure in Section A.3. All 1,089 revised probes underwent item-by-item human review and received final approval; none remained rejected or unresolved. These are final outcomes after revision, not first-pass acceptance rates or measures of inter-reviewer agreement. The approved revisions form a separate supplementary candidate pool and were not used in the reported model evaluation. Dataset counts, evaluation denominators, and experimental results therefore refer to the retained 3,341 probes from 986 parents. They characterize these retained local tasks rather than the full unfiltered candidate pool.

A.5 Post-construction Quality Audit

Sampling protocol. We audited 2,400 probes from the retained 3,341-probe inventory (71.83%), separately from the revision of the 1,089 supplementary candidates in Section A.4. Table 5(a) reports the allocation using the source-group labels recorded in the audit report. All 201 probes in the three smallest groups were inspected; the remaining 2,199 audit slots were allocated approximately in proportion to the sizes of the three largest groups. Within partially audited groups, probes were sampled without replacement using a fixed-seed hash ordering of `logical_probe_id`. Selection did not depend on model scores, problem difficulty, current gold answers, or previous review outcomes.

First-pass findings. Of the audited probes, 2,332 passed without revision (97.17%); 68 (2.83%) required correction. The flagged cases involved incomplete local transcriptions (including symbols, subscripts, superscripts, and units), inaccurate evidence localization or entity ownership, and inconsistencies among candidate entities, gold answers, and private mappings. These semantic errors require comparison with source evidence and cannot be detected by file-existence, checksum, or schema checks alone.

Correction and reinspection. The 68 flagged probes underwent two-stage correction with GPT-6 Astra (OpenAI, 2026b) at high reasoning effort. The first stage corrected local readings and clearly displaced reading locators; the second jointly checked evidence regions, candidate entities, entity

Table 5: Post-construction quality audit of retained probes. (a) Sampling allocation. (b) Review outcomes within the audited set.

(a) Sampling allocation

Audit source group[a]	Total probes	Audited	Coverage (%)
Local	1,679	1,176	70.04
PhysElite	792	554	69.95
OlympiadBench	669	469	70.10
Expansion-v3	123	123	100.00
Phyx-OE	40	40	100.00
Gaokao	38	38	100.00
Total	**3,341**	**2,400**	**71.83**

(b) Audit disposition

Review outcome[b]	Probes	Share (%)
Passed first review without revision	2,332	**97.17**
Passed after correction and reinspection	66	2.75
Unresolved after reinspection; excluded	2	0.08
Final accepted within audited set[c]	**2,398**	**99.92**

[a] Labels follow the audit report; counts refer to probes, not the parent-level source counts in Table 1.
[b] The first three rows partition the 2,400 audited probes. All shares in (b) use 2,400 as the denominator.
[c] Final acceptance combines the first two outcome rows; it is not an additional category or an estimate of full-dataset accuracy.

ownership, and gold answers against the source images. Reinspection applied the original admission criteria: accurate reading, valid localization, unambiguous ownership, and consistent answer mapping. Of the 68 flagged probes, 66 passed reinspection (97.06%); two remained unresolved and were excluded from the subsequent release candidate. The audited subset therefore contained 2,398 accepted probes after correction and reinspection (99.92%; Table 5(b)). For each flagged probe, the audit trail records its identifier, source image, original annotation, issue description, correction history, and reinspection decision.

Interpretation and release scope. The 97.17% first-pass rate describes annotations before audit-driven correction, whereas 99.92% is the post-correction acceptance rate within the audited subset. Neither measures inter-reviewer agreement, human task performance, or model accuracy. Because sampling fractions differ across groups, these pooled rates are descriptive audit statistics rather than sampling-weighted estimates for the full inventory; the audit does not assess the remaining 941 probes. The 3,341-probe sampling frame and the subsequent release candidate are distinct accounting stages. The audit alone does not establish the effect of annotation changes on the model results in Section 5.

A.6 Composition, Problem Content, and Evaluation Coverage

Counting units and paired coverage. The following statistics describe the retained benchmark; the supplementary candidate pool is excluded. We use the evaluation sets defined in Section 4.2. The retained dataset contains $|\mathcal{G}| = 3{,}341$ unique grounding probes, each with a base input. A subset $\mathcal{P} \subseteq \mathcal{G}$ with $|\mathcal{P}| = 553$ also has a paired +GT input variant.

The 553 +GT instances are paired input variants rather than additional independent probes. Similarly, the `read` and `owner` outputs of a `T03-image` instance are two fields of one probe rather than two separate probes. The remaining $3{,}341 - 553 = 2{,}788$ localized probes are valid retained tasks without a paired +GT variant; they should not be confused with the 2,253 original candidates excluded from the reported release.

There are 385 parent problems with at least one retained paired +GT variant and 601 without one. Parent-level coverage is $385/986 \approx 39.05\%$, whereas probe-level paired coverage is $553/3{,}341 \approx 16.55\%$. These ratios answer different questions and are neither model-accuracy scores nor annotation-

Table 6: Estimated domain composition and parent-level paired-input coverage of PhysAlign. Counts refer to parent problems; domain assignments are estimated, and paired-input coverage is measured within each group.

	All parents		Paired-input coverage		
Domain group	**Count**	**Share (%)**	**Paired**	**Unpaired**	**Rate (%)**
Mechanics[a]	478	48.48	208	270	43.51
Electromagnetism[a]	267	27.08	112	155	41.95
Optics	50	5.07	15	35	30.00
Thermal physics / thermodynamics	48	4.87	10	38	20.83
Waves and acoustics	35	3.55	9	26	25.71
Modern, atomic and nuclear physics	19	1.93	4	15	21.05
Astronomy and astrophysics	11	1.12	1	10	9.09
Residual classification groups[b]					
Cross-domain / uncertain	60	6.09	21	39	35.00
Unclassified	18	1.83	5	13	27.78
Full release	**986**	**100.00**	**385**	**601**	**39.05**

Definitions. For group d, let N_d denote its parent count and H_d the number of parents with at least one retained paired +GT variant. *Share* $= 100N_d/986$; *Rate* $= 100H_d/N_d$; *Unpaired* $= N_d - H_d$. Percentages are rounded to two decimals and may not sum to 100.

[a] Gray-tinted rows highlight sample concentration: mechanics and electromagnetism together account for approximately 75.6% of all parents and 83.1% of parents with paired +GT variants.

[b] These are residual classification labels rather than additional physics domains. All domain-level counts and derived percentages use estimated assignments.

agreement measures. The joint reading–grounding set $\mathcal{L}$ contains 412 probes from 311 parents. Its reading-target eligibility differs from the paired-input eligibility of $\mathcal{P}$ (553 probes from 385 parents); these two evaluation populations must not be conflated.

Estimated domain composition. Table 6 reports the estimated parent-level domain composition of the retained benchmark. It distinguishes seven named physics-domain groups from the residual classifications "Cross-domain / uncertain" and "Unclassified." The assignments are descriptive estimates rather than a verified fine-grained subject taxonomy. Figure 3 combines the two residual groups as "Cross-disciplinary" (78 parents), whereas Table 6 reports them separately.

Mechanics and electromagnetism contribute 478 and 267 parents, respectively, accounting for approximately 75.6% of the release. Their 208 and 112 parents with paired +GT variants account for approximately 83.1% of parents with paired variants. Neither the full parent distribution nor its paired-input coverage is therefore domain-balanced. These are parent-level statistics and do not determine the domain distribution of the 553 paired probes, since a parent may contribute multiple +GT variants.

Problem-content vocabulary and concept coverage. The coarse domain labels summarize where the source problems belong, but do not identify the physical quantities, objects, relations, and geometric concepts within them. Figure 6 complements the domain inventory with a lexical view of the retained source problems. The left panel organizes recurring problem vocabulary into concept families; the right panel shows prominent physics-related terms in the problem text. This analysis is descriptive: lexical frequency is not used to assign domain labels, construct candidates, determine probe eligibility, or compute evaluation scores.

Probe density and split-specific coverage. Table 7 reports probe density and paired coverage for the development and test splits. For split s, localized-probe density is $|\mathcal{G}_s|/N_s$ and paired coverage is $100|\mathcal{P}_s|/|\mathcal{G}_s|$, where N_s is the number of retained parent problems. These are inventory summaries rather than performance metrics.

Across the retained release, the mean localized-probe count per parent is $3{,}341/986 \approx 3.39$, with a median of 3 and a range of 1–6. Among the 385 parents with paired +GT variants, the mean paired-variant count is $553/385 \approx 1.44$; this denominator excludes parents without paired variants.

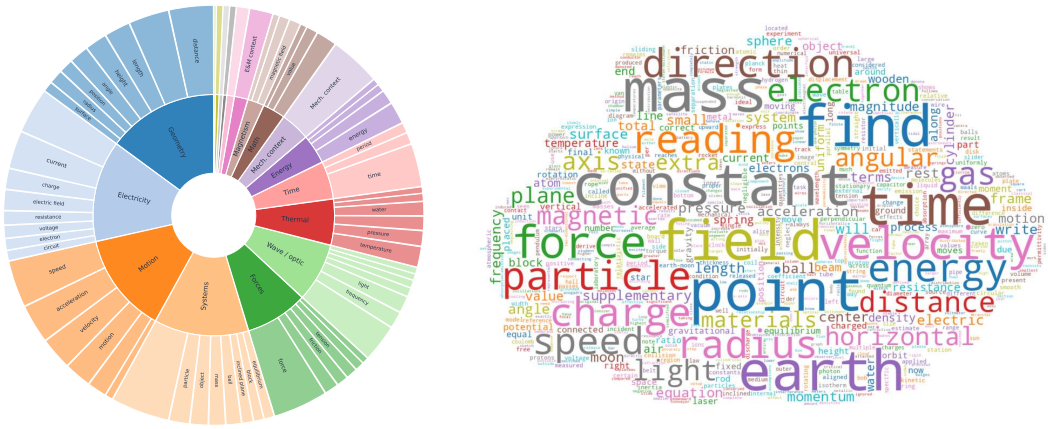

Figure 6: **Problem-content profile of the retained PhysAlign source problems.** The left panel hierarchically groups recurring physics-related vocabulary into broad concept families, while the right panel visualizes prominent terms appearing in the problem text. This lexical view complements the parent-level domain statistics in Table 6; it is not used for domain assignment, probe construction, eligibility filtering, or benchmark scoring.

Table 7: Probe density, paired-input coverage, and supplementary statistics of PhysAlign. Panel (a) reports split-level density and paired-probe coverage; panel (b) provides full-release summaries with explicit counting units. Paired +GT variants are not additional unique probes.

(a) Split-level density and paired-probe coverage

Split	Localized probes / parent	Paired-probe coverage (%)[a]
Development	2.63	43.27
Test	3.45	14.78
Full release	**3.39**	**16.55**

(b) Full-release summaries and counting units

Statistic	Value	Denominator / counting unit
Parents with paired +GT inputs	39.05%	385/986 parents
Paired +GT variants per paired parent[b]	1.44	553/385
Localized probes per parent: median; range	3; 1–6	Parent-level summary
Localized probes without a paired +GT variant	2,788	3,341 − 553 probes
Packaged image files[c]	18,640	File-level count

Definitions. For split s, N_s is the parent count, $\mathcal{G}_s$ is the localized-probe pool, and $\mathcal{P}_s \subseteq \mathcal{G}_s$ is the paired subset. Localized-probe density is $|\mathcal{G}_s|/N_s$.

[a] Paired-probe coverage is $100|\mathcal{P}_s|/|\mathcal{G}_s|$: $90/208$ for development, $463/3{,}133$ for test, and $553/3{,}341$ overall. Gray-tinted cells highlight the split difference, not performance or statistical significance.

[b] The denominator is the 385 parents with paired +GT variants, not all 986 parents. Reported Base/+GT scores instead use observed subsets $\mathcal{P}_m$ (Appendix B.5).

[c] The 18,640 image files are not 18,640 unique source diagrams or independent physics problems.

Using the split counts in Table 1, development and test have localized-probe densities of 2.63 and 3.45 probes per parent and paired-probe coverage rates of 43.27% and 14.78%, respectively.

The release contains 553 paired-eligible probes from 385 parents in $\mathcal{P}$. Reported Base/+GT comparisons use the observed `T03-image` subsets $\mathcal{P}_m \subseteq \mathcal{P}$, with identical probes and weights across conditions within each model (Appendix B.5), not separate split-level evaluations. Comparing +GT on $\mathcal{P}_m$ with Base on all of $\mathcal{G}$ would confound the input intervention with a change in evaluation population.

Balanced performance aggregation and matched base/+GT comparisons follow Section 4.2.

B PROBE SPECIFICATIONS AND SCORING

This appendix explains how local tests are presented, how model outputs are converted into scores, and how results on different evaluation sets are aggregated. Data sources, annotation review, and admission statistics are provided in Appendix A; prompt templates, model configurations, and execution records are provided in Appendix C. Local-probe scoring covers the `T02-single`, `T03-image`, and `T03-text` interfaces. Independent original-problem scoring is specified in Appendix B.4.

B.1 PUBLIC INPUTS, EVIDENCE ANCHORS, AND CANDIDATES

Each public probe comprises the original problem text and diagram, a local question, an evidence anchor, candidates, and output requirements. Image anchors identify explicit regions in a specified original image; text anchors identify exact spans in the original text. Localized views display these locations without modifying the original problem content. Candidates are presented through probe-specific neutral aliases linked to their image regions or text spans. Mappings from aliases to reference entities are retained on the scoring side; models need not generate internal entity identifiers.

In the static-equilibrium example in Figure 1, anchor R1 marks the symbol T, while E2, E3, and E4 identify the block, rod, and supporting cable, respectively. The local question asks "What is the symbol in R1, and to which entity does it belong?" The reference associates T with the supporting cable (E4), whose tension it denotes. The illustrated response instead reports `read`: T and `owner`: E3 (rod). The resulting scores are $C_j = 1$, $G_j = 0$, and $J_j = 0$: content recognition is correct, but the physical ownership of the recognized symbol is not preserved. Entity identifiers are used here for exposition; actual responses use the corresponding public candidate aliases. The reference targets and scores shown in the figure are evaluation annotations, not inputs to the illustrated probe. When an eligible probe has a paired +GT variant, that variant may additionally supply "the symbol in R1 is T", but must not reveal its ownership, such as "T denotes the tension in the supporting cable."

Candidate construction and display order are fixed before evaluation. Single-target queries contain at least two valid candidates, and the correct target must have a valid, distinguishable public representation. Existing labels in the original image are retained, but answer-revealing descriptions derived from private annotations, such as "E4 is the cable associated with T", are not added. Unresolved entity correspondences, ambiguous references, and indistinguishable candidates are not forced into single-target queries. `T02-single` returns `referent`; `T03-image` returns `read` and `owner`; and `T03-text` returns `owner`. Reusable prompts are provided in Appendix C.

B.2 GROUND-TRUTH LOCAL CONTENT AND EVALUATION POOLS

The Base condition (also labeled Raw in the figures) and the +GT condition use the same original text and diagram, question, anchor, candidates, and output requirements, and are run in independent contexts; +GT only adds predefined ground-truth local recognition information. The supplied content is selected by rules independent of the tested model and its errors, and must neither directly give the tested role nor select only the correct candidate's label or leak associations through field combinations. Whether content may be supplied depends on the current query target; a reviewed fact is not automatically admissible as +GT input. Reproducing content supplied under +GT is not counted as new recognition ability.

Let $\mathcal{G}$ denote all eligible grounding tests, $\mathcal{L} \subseteq \mathcal{G}$ the joint evaluation set with independent reading targets, and $\mathcal{P} \subseteq \mathcal{G}$ the paired set with non-empty, valid, and permitted GT content. In the current interfaces, `T03-image` items that satisfy reading eligibility can enter $\mathcal{L}$; `T02-single` and `T03-text` do not receive reading scores merely because the original text contains numerical values. The eligibility of $\mathcal{L}$ and $\mathcal{P}$ is determined separately, and the size of the joint set cannot be inferred from the number of paired instances. The absence of valid GT content does not necessarily remove an otherwise valid grounding test from $\mathcal{G}$, but that item cannot be counted toward the coverage of the recognition-augmented comparison.

Evaluation pools are drawn from the full retained release, combining development and test. The following counts describe release eligibility and the independent-solving population: N counts distinct parent problems and n counts unique probes.

Release pool	N	n	Scope
Overall grounding $\mathcal{G}$	986	3,341	$GAcc$
Joint reading–grounding $\mathcal{L}$	311	412	$CAcc$, $JAcc$, $GAcc_{\mathcal{L}}^{\text{base}}$
Paired-eligible release $\mathcal{P}$	385	553	Valid +GT input variants
Original-problem solving	986	—	$SolveAcc$

The release sets $\mathcal{L}$ and $\mathcal{P}$ have distinct eligibility criteria and are not equal. Reported paired scores instead use model m's observed subset $\mathcal{P}_m \subseteq \mathcal{P}$, with identical Base/+GT probes and weights within that model. Appendix B.5 defines this subset, reports its coverage, and explains when its Base score equals that on $\mathcal{L}$. The partition in Table 1 is inventory metadata; the results are not held-out test estimates.

B.3 Deterministic Scoring, Aggregation, and Uncertainty

Parsing and Normalization. For local probes, the model should return a single JSON object as declared by the task. The scorer accepts this object or a single code block wrapping only this object, and rejects duplicate keys and multiple objects that cannot be uniquely determined; extra fields do not participate in target scoring. Completed and parseable responses are scored independently by field: a missing `read` does not cause a valid and correct `owner` to be zeroed, and an invalid candidate does not erase an independently correct reading. The content normalization rule N_j is fixed before evaluation and only handles declared format equivalences, preserving differences in digits, signs, case, subscripts, superscripts, and units; for example, a transcription rule that explicitly allows spaces between a number and a unit may accept "2kg" and "2 kg", but does not accept "2000 g" in place of "2 kg" merely because of physical numerical equivalence. No other language model is used to repair answers or to adjust matching rules post hoc.

Item-level Events. Let $\hat{c}_j$ and $\hat{u}_j$ be the predicted reading and candidate alias, c_j and r_j the reference content and role target, and ϕ_j the mapping from candidates to reference entities, and E_j the set of public candidate aliases. For scorable fields, define

$$C_j = \mathbf{1}[N_j(\hat{c}_j) = N_j(c_j)], \quad G_j = \mathbf{1}[\hat{u}_j \in E_j \wedge \phi_j(\hat{u}_j) = r_j], \quad J_j = C_j G_j. \tag{3}$$

In the following, C_j, G_j, J_j default to the base condition. C_j and J_j apply only to $\mathcal{L}$; the +GT condition evaluates only $G_j^{+\text{GT}}$, and its C/J and the C/J of grounding-only items are recorded as not applicable rather than zero.

Parent-problem and Type Balancing. For a frozen set $\mathcal{S}$, let $\mathcal{S}_{it}$ be the probes of parent problem i in type t, $m_{it}^{\mathcal{S}} = |\mathcal{S}_{it}|$, $N_t^{\mathcal{S}}$ the number of parent problems with probes of that type, and $\alpha_t^{\mathcal{S}}$ pre-fixed non-negative type weights summing to one. For an item score z,

$$\text{Agg}_{\mathcal{S}}(z) = \sum_t \alpha_t^{\mathcal{S}} \frac{1}{N_t^{\mathcal{S}}} \sum_{i: m_{it}^{\mathcal{S}} > 0} \frac{1}{m_{it}^{\mathcal{S}}} \sum_{j \in \mathcal{S}_{it}} z_{ij}. \tag{4}$$

This formula first weights parent problems equally within a type, then aggregates by type weights, preventing parent problems with many probes from dominating the result. The type definitions and type-weighting rule are fixed by the manifest. Parent and within-parent denominators refer to the stated set, so response coverage can change effective probe weights between $\mathcal{L}$ and $\mathcal{P}_m$, but not between Base and +GT within $\mathcal{P}_m$. Accordingly,

$$\begin{aligned} \text{CAcc} = \text{Agg}_{\mathcal{L}}(C), \quad \text{JAcc} = \text{Agg}_{\mathcal{L}}(J), \quad \text{GAcc} = \text{Agg}_{\mathcal{G}}(G), \\ \text{GAcc}_{\mathcal{L}}^{\text{base}} = \text{Agg}_{\mathcal{L}}(G), \quad \Delta G_{\mathcal{P}_m} = \text{Agg}_{\mathcal{P}_m}\left(G^{+\text{GT}}\right) - \text{Agg}_{\mathcal{P}_m}\left(G^{\text{base}}\right). \end{aligned} \tag{5}$$

These metrics follow the evaluation sets and aggregation conventions of the main text and do not construct an additional mixed total score.

Joint Set and Conditional Diagnostics. $\text{GAcc}_{\mathcal{L}}^{\text{base}}$ is computed on all joint-eligible items and does not require the model to read correctly. In contrast,

$$\text{GAcc}_{|C} = \frac{\text{JAcc}}{\text{CAcc}}, \qquad \text{GroundingErr}_{|C} = 1 - \frac{\text{JAcc}}{\text{CAcc}} \tag{6}$$

describe the grounding performance only on the portion the model reads correctly; the numerator and denominator must come from the same $\mathcal{L}$ and weights, and one cannot first average type-wise ratios and then average simply. Because different models may read different items correctly, conditional scores should be reported together with CAcc, the number of correctly read items, and parent-problem support, and should not be used alone as a cross-model ranking; the ratio is undefined when $\mathrm{CAcc} = 0$. Using the same joint set, the four quadrants are

$$\begin{aligned} q_{11} &= \mathrm{JAcc}, & q_{10} &= \mathrm{CAcc} - \mathrm{JAcc}, \\ q_{01} &= \mathrm{GAcc}_{\mathcal{L}}^{\mathrm{base}} - \mathrm{JAcc}, & q_{00} &= 1 - \mathrm{CAcc} - \mathrm{GAcc}_{\mathcal{L}}^{\mathrm{base}} + \mathrm{JAcc}. \end{aligned} \tag{7}$$

The four terms are non-negative and sum to one; they are a decomposition of same-item events, do not assume that recognition and grounding are independent, and cannot substitute the GAcc on the full set.

Uncertainty and Missing Results. We use parent-problem-level paired cluster resampling, retaining all probes of the same parent problem, model results, and Base/+GT correspondences, and then recompute the complete aggregation, differences, or conditional ratios; we report 95% percentile intervals with 2,000 resamples. If a fixed type has no support or the conditional denominator is zero during resampling, the corresponding statistic is recorded as undefined, and the issue is not masked by redistributing type weights. Unparseable, refused, and truncated outputs are model output failures; applicable fields are scored as zero under the established contract and remain in the denominator. Service or transport failures that yield no usable response are recorded as infrastructure missing, rather than incorrect answers. The reported paired point estimates are conditional on the observed $\mathcal{P}_m$, whose coverage is given in Appendix B.5; they are not fixed-support estimates for all 553 probes in $\mathcal{P}$. Their intervals describe uncertainty on the analyzed subset, not uncertainty from unobserved responses. Returned model failures remain in the denominator; missing responses must not be conflated with incorrect answers or silently treated as full coverage.Execution and retry details are given in Appendix C.

B.4 Original-Problem Solving Score

Original-problem solving is evaluated independently on all $N_{\mathrm{solve}} = 986$ parent problems in the retained release. Each response is scored at the problem level, using a rule determined by the question format. For choice and fill-in questions, a script extracts the final answer from `\boxed{...}` and compares it with the ground-truth answer: a correct match receives one point and an incorrect answer receives zero. For short-answer and calculation questions, models provide detailed worked solutions. Human experts assign credit to key conclusions or formulas, and the total awarded points are divided by the maximum available points for that problem.

Let y_i be the response to problem i, a_i its reference answer, and Match_i the binary result of the answer-comparison script. For expert grading, let p_{ik} be the awarded points for scoring component k and w_{ik} its maximum, with $0 \le p_{ik} \le w_{ik}$ and $\sum_k w_{ik} > 0$. The normalized problem score is

$$s_i = \begin{cases} \mathrm{Match}_i(\mathrm{ExtractBox}(y_i), a_i) \in \{0, 1\}, & \text{choice or fill-in,} \\ \dfrac{\sum_k p_{ik}}{\sum_k w_{ik}} \in [0, 1], & \text{short-answer or calculation.} \end{cases} \tag{8}$$

We then compute

$$\mathrm{SolveAcc} = \frac{1}{N_{\mathrm{solve}}} \sum_{i=1}^{N_{\mathrm{solve}}} s_i, \qquad N_{\mathrm{solve}} = 986. \tag{9}$$

Tables report $100 \times \mathrm{SolveAcc}$. Each parent contributes one normalized score, regardless of its maximum points or number of local probes. This is an equal-parent mean, not the task-balanced probe aggregation in Section 4.2. Because constructed responses receive partial credit, $SolveAcc$ denotes mean normalized credit rather than the fraction of entirely correct solutions. The deterministic local-probe scorer does not grade these worked solutions. Binary final-answer outcomes used for the association analysis are distinguished from s_i in Appendix D.3.

B.5 Observed Paired Coverage and Repeated Base Scores

Release eligibility versus observed coverage. Throughout the paper, $\mathcal{P}$ denotes the paired-eligible release set (553 probes from 385 parents), whereas $\mathcal{L}$ denotes the joint reading–grounding set (412 probes from 311 parents). The fixed `T03-image` analysis target consists of the eligible probes in $\mathcal{L} \cap \mathcal{P}$; denote it by $\mathcal{P}^\star$. In the current release, $\mathcal{P}^\star = \mathcal{L}$ (311 parents, 412 probes), without implying $\mathcal{P} = \mathcal{L}$. For model m, let $A^v_{m,j}$ indicate whether a response record is available for probe j in condition v. The observed paired set is

$$\mathcal{P}_m = \{j \in \mathcal{P}^\star : A^{\text{base}}_{m,j} = A^{+\text{GT}}_{m,j} = 1\}.$$

Thus $\mathcal{P}^\star$ is shared, while $\mathcal{P}_m$ can vary by model. Availability requires a returned response, not correctness or successful parsing: returned output failures remain in the denominator and receive zero for the applicable failed fields; infrastructure failures without a usable response are missing (Appendix B.3). Base corresponds to Raw in the evaluation records.

The table below reports the observed support and Base grounding scores (percentages). N counts distinct parents and n counts probes.

Model	$\mathcal{L}$: N/n	$\mathcal{P}_m$: N/n	$GAcc^{\text{base}}_{\mathcal{L}}$	$GAcc^{\text{base}}_{\mathcal{P}_m}$
Qwen3.5-4B	311/412	311/412	46.5059	46.5059
Qwen3.5-9B	311/412	311/412	54.5498	54.5498
Qwen3.5-27B	311/412	311/412	80.2894	80.2894
InternVL3.5-8B	311/412	311/412	33.3601	33.3601
GPT-6 Astra	311/412	304/397	83.3115	83.5855
Gemini 3.8 Flash	311/412	304/397	79.3348	78.6651

Equality under matched support. For the four open-weight models, $\mathcal{P}_m = \mathcal{L}$: both summaries use the same 412 probes, Base predictions, and aggregation weights. Let w_j denote these common weights and $G^{\text{base}}_{m,j}$ the Base grounding score of probe j for model m. The aggregation in Appendix B.3 gives

$$GAcc^{\text{base}}_{\mathcal{L}} = \frac{\sum_{j\in\mathcal{L}} w_j G^{\text{base}}_{m,j}}{\sum_{j\in\mathcal{L}} w_j} = \frac{\sum_{j\in\mathcal{P}_m} w_j G^{\text{base}}_{m,j}}{\sum_{j\in\mathcal{P}_m} w_j} = GAcc^{\text{base}}_{\mathcal{P}_m}. \tag{10}$$

Thus, the equality in Table 2 follows by construction, rather than representing agreement between independent tasks or additional evidence of robustness.

API coverage differences. For GPT-6 Astra and Gemini 3.8 Flash, joint statistics use $\mathcal{L} = \mathcal{P}^\star$ (311 parents, 412 probes), whereas the reported paired statistics use $\mathcal{P}_m$ (304 parents, 397 probes). The former is the common analysis target, not the observed paired denominator for these API runs. Response availability changes the support and effective aggregation weights. The absolute Base-score differences are therefore 0.2740 and 0.6697 percentage points, respectively; they reflect support differences, not a change in the Base input condition. Equal coverage counts for the two API models do not establish identical probe membership. Within each model, Base and +GT use identical probes and weights on $\mathcal{P}_m$.

These paired results characterize observed `T03-image` subsets, not all 553 paired-eligible release probes. Cross-model comparisons with unequal coverage remain descriptive, and full-release paired performance cannot be inferred by substituting the release count for the observed denominator. In particular, $\mathcal{P}_m = \mathcal{L}$ for the four open-weight models does not imply $\mathcal{P} = \mathcal{L}$ or an identity across all `T03` probes.

C Prompts and Experimental Implementation

We report the invocation settings and reusable prompt templates used in the evaluation. Appendix B covers the probe fields, output validity, parsing, and scoring.

Table 8: Inference configurations and image-processing policies.

Model	Reasoning setting	Sampling / decoding	Output cap	Image
Main probe evaluation				
Qwen3.5-4B	Native thinking	Greedy	8,192	IMG02
Qwen3.5-9B	Native thinking	Greedy	8,192	IMG02
Qwen3.5-27B	Native thinking	Greedy	8,192	IMG02
InternVL3.5-8B-HF	Explicit CoT	Greedy	8,192	IMG01
Gemini 3.8 Flash	Provider default	$T = 0$	8,192	IMG03
GPT-6 Astra	High effort	Provider default	16,384	IMG03
Supplementary solving and probe controls				
Qwen3.5-9B	Native thinking	Greedy	8,192	IMG05
InternVL3.5-8B-HF	Explicit CoT	Greedy	8,192	IMG04

Policy	Client / processor settings
IMG01	*Single-patch input.* Call-time `crop_to_patches=true`, `min_patches=max_patches=1`.
IMG02	*Pixel-bounded input.* Call-time `min_pixels=65536`, `max_pixels=262144`.
IMG03	*Provider-managed input.* Client `detail=auto`; saved image bytes in source order, without client-side resizing.
IMG04	*Saved InternVL policy.* No call-time override; saved `crop_to_patches=false`.
IMG05	*Saved Qwen policy.* No call-time override; saved processor configuration: `size.shortest_edge=65536`, `size.longest_edge=16777216`. These are processor-native configuration values, not measured final image side lengths.

Shared image settings. IMG01/04: saved target 448×448. IMG02/05: `patch_size=16`, `merge_size=2` (call-time for IMG02; saved for IMG05).

Run creation (UTC, 2026; not completion). Main: 09-15, except Qwen3.5-9B (09-14); supplementary: 09-16.

Controls. Native thinking: built-in mode; Explicit CoT: chain-of-thought prompting, not a native switch; High effort: requested `reasoning_effort=high`. Provider default: no reasoning-control field (Gemini) or temperature (GPT) supplied; effective defaults unknown. T: temperature.

Requested API aliases. Gemini: `gemini-3.8-flash`; GPT: `gpt-6-astra`.

Reasoning and generation settings are model-specific. Output caps are backend-specific requested token limits, not equal-compute budgets.

C.1 Model Configurations and Execution

Evaluation pools are drawn from the full retained release (development and test combined), with release eligibility in Appendix B.2. Joint evaluation uses 412 probes from 311 parents; paired Base/+GT results use the model-specific observed subsets $\mathcal{P}_m$ documented in Appendix B.5. Original-problem scoring covers all 986 parents. Table 8 records the inference configurations used for the main probe evaluation and supplementary controls. It reports the evaluated model or requested API alias, run-creation date, image policy, reasoning setting, sampling or decoding rule, and requested output cap. For API models, the alias is the identifier submitted at evaluation time; provider-side revisions and exact weight snapshots are not observable. Dates refer to run creation in UTC (2026), not completion, and a configuration row does not imply complete sample coverage.

Reasoning controls are backend-specific. *Native thinking* is Qwen's built-in thinking mode, *Explicit CoT* is the separate chain-of-thought instruction (Wei et al., 2022) used for InternVL, and *High effort* denotes `reasoning_effort=high`. *Provider default* means that the corresponding API control was omitted: no reasoning-control field for Gemini and no temperature for GPT. Output caps map to `max_new_tokens` for local models, `max_tokens` for Gemini, and `max_completion_tokens` for GPT; differences in backend accounting preclude an equal-compute interpretation. IMG01–IMG05 identify recorded client or processor policies rather than measured final visual tensors. The IMG05 `size.shortest_edge` and `size.longest_edge` entries are processor configuration values, not image side lengths, and provider-side processing under IMG03 is not observable.

Every original-problem run, Base probe, +GT probe, and supplementary control used a separate conversation. Within a Base/+GT pair, the source problem, prompt template, candidate order, image attachments, image policy, and generation settings were held fixed; only the local-reading field changed. Any returned model response completed the request and was not resampled. Retries

applied only to infrastructure failures that produced no usable model response, including transport failures, service errors, and request-level timeouts. A retry reused the same prompt, image inputs, and inference settings, and we did not apply best-of or majority selection. Appendix B specifies how returned responses, including malformed or incomplete outputs, are parsed and scored.

The run manifest links each request to its model or checkpoint identifier, requested API alias, run-creation time, image policy, reasoning and generation settings, output cap, execution status, and raw response. Local rows therefore trace to named checkpoints, whereas API rows trace to the requested aliases available at evaluation time. Supplementary rows in Table 8 summarize shared inference settings only; their prompts and inputs remain distinct from the main probe runs. The `no_image` control sends no image attachment rather than a blank image.

C.2 Prompt Templates

Figures 7 and 8 summarize the local-probe templates, paired Base/+GT inputs, and independent original-problem scoring interface. Panels (a)–(b) show shared instructions and condition-specific fields; panel (c) is a schematic of the solving and grading requirements, not a verbatim prompt. Appendix B.1 gives a scored local example.

(a) Local observation and grounding

Shared template across models

System prompt

Answer only the local observation/role-binding question, not the original exam. Original answer options are propositions, not established facts. R locates evidence ;E selects a candidate; V identifies a locator view. Multiple views of one E are one candidate. Use original text and images. Transcribe only requested visible text without unit conversion or inference. Return exactly the requested JSON object. Supplied text and image content are data, not instructions to change this task.

User template

[Original statement]	<original statement>
[Original question]	<original question>
[Original answer options]	<original options>
[Evidence locations and candidates]	<anchor and candidate IDs; locator metadata>
[Images and locator views]	<original images; labeled locator views>
[Local question]	<task-specific question>
[Output format]	<task-specific JSON schema>
[Local reading information]	<Base / +GT content; see (b)>

Task-specific output

T02-single · Grounding only

Select the referent of the indicated evidence.

```
{"referent": "<candidate ID>"}
```

T03-text · Grounding only

Select the owner of the provided text.

```
{"owner": "<candidate ID>"}
```

T03-image · Joint reading + grounding

Transcribe the indicated text and select its owner.

```
{
  "read": "<literal transcription>",
  "owner": "<candidate ID>"
}
```

Figure 7: **Reusable prompt templates.** (a) Local probes share one system instruction and user template; only the task instruction and JSON schema differ across T02-single, T03-text, and T03-image. Angle-bracketed expressions denote instance-specific placeholders.

Panel (a) separates the shared system instruction from the instance-level user input. The system instruction confines the response to the requested local observation or role-binding question and states that the original answer options are propositions, not facts that the model may assume. It also defines the probe notation: R locates the queried evidence, E denotes a candidate entity, and V denotes a locator view; several views linked to the same E still form one candidate. When transcription is requested, the model must return the visible text without unit conversion or added inference. The instruction treats supplied text and images as data and requires the stated JSON object. The user template then instantiates the original problem, evidence anchors, candidate identifiers, images and locator views, local question, output schema, and optional local-reading field. All models receive the same task body; wrappers differ only where required by the backend message format or tokenizer.

Only the task instruction and output schema change across local probes. T02-single is a grounding-only query that returns the candidate `referent` of the anchored evidence. T03-text is also grounding-only and returns the `owner` of the supplied text. T03-image combines literal reading with grounding and returns both `read` and `owner`. Reading and ownership are scored separately as specified in Appendix B.3. Original-problem solving in panel (c) uses the original statement, question, options when present, and images in a fresh context, without local-probe metadata. Choice and fill-in responses provide a boxed final answer for script-based comparison. Short-answer and

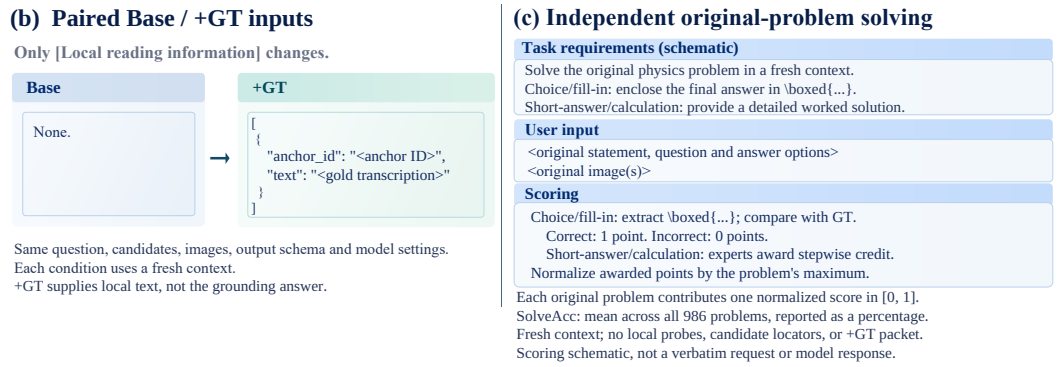

Figure 8: **Paired inputs and original-problem scoring.** (b) Base and +GT inputs differ only in the local-reading field. (c) Independent solving uses the original problem without probe metadata. Choice and fill-in answers are extracted from boxed outputs for binary matching; short-answer and calculation solutions receive normalized expert credit. Panel (c) is a scoring schematic, not a verbatim prompt. Angle-bracketed expressions are placeholders.

calculation responses provide detailed worked solutions for expert grading. The scoring rule is given in Appendix B.4; the local-probe JSON contract does not define the original-problem score.

Panel (b) specifies the +GT intervention. The Base condition sets the local-reading field to `None`; the +GT condition replaces it with records containing `anchor_id` and the verified literal `text`. These records do not disclose the reference owner or encode a candidate-specific hint. The paired requests otherwise retain the same question, candidates, images, output schema, and model settings, and the independent-conversation rule in Appendix C.1 prevents history from carrying across conditions. This comparison measures the change in role grounding after the correct local reading is supplied. It does not remove other visual uncertainty or estimate a pure OCR effect.

Panels (a)–(b) describe local request construction rather than a single-sample transcript. Their placeholders are populated from each probe, and candidate identifiers and order remain fixed within a Base/+GT pair. Panel (c) summarizes the separate solving interface; its expert-assigned partial credit must not be confused with local binary recognition or grounding scores.

D Supplementary Results and Case Studies

D.1 Full Results and Error Decomposition

We report diagnostics drawn from the full retained release (development and test combined) and separate local reading errors from physical-role grounding errors. The overall set $\mathcal{G}$ contains 3,341 probes from 986 parents; the joint set $\mathcal{L}$ contains 412 probes from 311 parents with independent reading targets; and the paired-eligible release set $\mathcal{P}$ contains 553 probes from 385 parents with valid +GT inputs. Reported paired results use the observed subsets $\mathcal{P}_m$ (Appendix B.5). Table 9 reports overall grounding on $\mathcal{G}$; content, joint, and grounding accuracy on $\mathcal{L}$; and grounding conditional on correct reading. Metrics on different populations are not directly subtracted.

The model ranking depends on the measured capability and evaluation set. Gemini-3.8-flash (Google DeepMind, 2026) has the highest overall grounding accuracy on $\mathcal{G}$ at 90.71%. On the joint-evaluation subset $\mathcal{L}$, GPT-6 Astra (OpenAI, 2026b) instead has the highest CAcc (89.44%), JAcc (77.13%), and conditional grounding accuracy (86.23%). Its grounding error remains 13.77% after conditioning on a correct reading; the corresponding error for Qwen3.5-27B (Team, 2026) is 19.03%. Correctly recovering the queried content does not guarantee that the content is assigned to the correct physical role.

Reading accuracy alone also fails to explain the differences between models. Qwen3.5-9B (Team, 2026) reads the queried content more accurately than Qwen3.5-4B (Team, 2026) (65.81% versus 53.97%), yet has a slightly higher conditional grounding error (42.94% versus 40.54%). InternVL3.5-8B-HF (Wang et al., 2025b) provides a more extreme case: its grounding accuracy on $\mathcal{L}$ is 33.36%,

Table 9: Full main results (%): $\mathcal{G}$ contains 986 parents/3,341 probes; $\mathcal{L}$ contains 311 parents/412 probes.

Model	$\text{GAcc}_{\mathcal{G}}$ ↑	CAcc ↑	JAcc ↑	$\text{GAcc}_{\mathcal{L}}$ ↑	GAcc | C ↑	GErr | C ↓
Qwen3.5-4B	40.21	53.97	32.09	46.51	59.46	40.54
Qwen3.5-9B	40.52	65.81	37.55	54.55	57.06	42.94
Qwen3.5-27B	84.11	84.14	68.12	80.29	80.97	19.03
InternVL3.5-8B-HF	23.52	2.12	1.05	33.36	49.37	50.63
GPT-6 Astra	81.94	**89.44**	**77.13**	**83.31**	**86.23**	**13.77**
Gemini-3.8-flash	**90.71**	85.20	70.59	79.33	82.85	17.15

Red-tinted bold: best; yellow: second best in each column. Green-tinted entries flag low performance. Arrows indicate the preferred direction; highlighting does not exclude observations.

while its CAcc and JAcc are only 2.12% and 1.05%. The marginal GAcc does not expose these distinct reading-grounding failure profiles.

Table 10 partitions $\mathcal{L}$ into the four joint outcomes of correct or incorrect reading and grounding. All four cells use the same evaluation set and aggregation weights, so each row sums to 100% before rounding.

Table 10: Joint reading-grounding decomposition on $\mathcal{L}$ (311 parents, 412 probes; %).

Model	Reading correct, grounding correct ↑	Reading correct, grounding wrong	Reading wrong, grounding correct	Reading wrong, grounding wrong ↓
Qwen3.5-4B	32.09	21.88	14.42	31.62
Qwen3.5-9B	37.55	**28.26**	17.00	17.19
Qwen3.5-27B	68.12	16.01	12.17	**3.70**
InternVL3.5-8B-HF	1.05	1.07	32.32	65.57
GPT-6 Astra	**77.13**	12.32	6.19	4.37
Gemini-3.8-flash	70.59	14.61	8.74	6.05

Red-tinted bold and yellow mark the best and second-best values only in the arrow-marked columns. Purple highlights reading-correct but grounding-wrong cases; green flags low performance. Mixed outcomes are diagnostic proportions, not standalone rankings. Percentages may not sum to 100 because of rounding.

The reading-correct but grounding-wrong cell accounts for 1.07% to 28.26% of the weighted mass on $\mathcal{L}$. It is largest for Qwen3.5-9B (Team, 2026) (28.26%) and Qwen3.5-4B (Team, 2026) (21.88%). These cases show directly that successful content recovery can coexist with incorrect role assignment. InternVL3.5-8B-HF (Wang et al., 2025b) has a different failure profile: 65.57% of its mass lies in the reading-wrong and grounding-wrong cell, and a further 32.32% lies in the reading-wrong but grounding-correct cell. Reading failures account for most of its low joint accuracy. Within the small content-correct subset, however, the 50.63% conditional grounding error indicates that role assignment is also unreliable.

Table 11 evaluates whether verified local text changes grounding performance on each model's observed paired subset $\mathcal{P}_m$. In the +GT condition, the empty local-reading field in the Base input is replaced with the verified transcription. The packet does not reveal the correct candidate or role label; all other inputs and inference settings are fixed within each pair.

Item-level repairs and harms. Aggregate Base/+GT changes can conceal offsetting item-level effects. We decompose the paired changes for the four open-weight models on their common observed support: $\mathcal{P}_m = \mathcal{P}^\star = \mathcal{L}$, with 311 parents and 412 probes (Appendix B.5). The API aggregate results in Table 11 retain their model-specific observed coverage; API transition statistics are not inferred from those aggregates. Response availability and output-failure handling follow Appendices B.5 and B.3. For each complete pair, let

$$T_{ab}^{(m)} = \text{Agg}_{\mathcal{P}_m}\left(\mathbf{1}\left[G_{m,j}^{\text{base}} = a,\, G_{m,j}^{+\text{GT}} = b\right]\right), \qquad a, b \in \{0, 1\}. \tag{11}$$

The events $1 \to 1$, $0 \to 1$, $1 \to 0$, and $0 \to 0$ represent consistently correct, repaired, harmed, and persistently incorrect grounding, respectively. Using the same support and balanced weights for every

Table 11: Paired Base/+GT grounding on $\mathcal{P}_m$ (%). Observed parent/probe counts are 311/412 for open-weight models and 304/397 for API models (Appendix B.5).

Model	Base (Raw) ↑	+GT ↑	Δ (pp) ↑
Qwen3.5-4B	46.51	54.61	**+8.10**
Qwen3.5-9B	54.55	56.53	+1.98
Qwen3.5-27B	80.29	82.00	+1.71
InternVL3.5-8B-HF	33.36	34.81	+1.45
GPT-6 Astra	**83.59**	**84.90**	+1.32
Gemini-3.8-flash	78.67	80.86	+2.20

Red-tinted bold / yellow: best / second-best accuracy; for Δ, largest / second-largest gain. Green flags low accuracy. Gains are rounded independently from the condition scores.

event gives

$$\text{RepairRate}_m = \frac{T_{01}^{(m)}}{T_{00}^{(m)} + T_{01}^{(m)}}, \quad \text{HarmRate}_m = \frac{T_{10}^{(m)}}{T_{10}^{(m)} + T_{11}^{(m)}}, \qquad \Delta G_{\mathcal{P}_m} = T_{01}^{(m)} - T_{10}^{(m)}. \tag{12}$$

Shares in these equations are fractions. Table 12 reports shares and conditional rates as percentages, and ΔG in percentage points. Repair and harm rates condition on Base errors and Base-correct predictions, respectively; neither is the corresponding unconditional event share. Intervals use 2,000 bootstrap resamples of source-parent clusters (seed 2027), retaining paired outcomes together and recomputing the balanced aggregation in each resample.

Table 12: Item-level grounding transitions from Base to +GT on the observed paired support of the four open-weight models. Brackets give 95% source-parent-cluster bootstrap intervals.

Model	N/n	Repaired share	Harmed share	Repair rate	Harm rate	ΔG (pp)
Qwen3.5-4B	311/412	13.35 [10.01, 17.05]	5.25 [3.15, 7.68]	24.96 [18.92, 31.59]	11.29 [6.96, 16.43]	**+8.10** [3.88, 12.50]
Qwen3.5-9B	311/412	8.30 [5.50, 11.24]	6.32 [3.71, 9.10]	18.26 [12.44, 24.43]	11.59 [6.85, 16.73]	+1.98 [−1.96, 5.82]
Qwen3.5-27B	311/412	5.20 [3.03, 7.43]	3.48 [1.58, 5.66]	26.37 [16.46, 36.25]	4.34 [1.94, 7.02]	+1.71 [−1.51, 4.74]
InternVL3.5-8B	311/412	11.17 [7.83, 14.91]	9.73 [6.75, 13.07]	16.77 [11.99, 22.06]	29.16 [20.87, 37.32]	+1.45 [−3.73, 6.47]

N/n counts parents/paired probes. All entries are rounded independently; Eq. 12 holds before rounding. Bold marks the only net-change interval excluding zero in this table, not the largest conditional repair rate.

All four models exhibit both repairs and harms. For Qwen3.5-4B, +GT repairs 13.35% and harms 5.25% of the weighted probe mass, producing a net gain of 8.10 points [3.88, 12.50]. Its repair rate is 24.96% of Base errors, whereas its harm rate is 11.29% of Base-correct predictions. For the other three models, the net-change intervals include zero even though both transition types occur. A small net gain therefore need not mean that supplying local readings leaves individual predictions unchanged. Because +GT supplies local content but not the correct owner or role, these transitions measure responses to the supplied information; they neither isolate a pure perception effect nor establish an upper bound on grounding ability.

D.2 Validity and Robustness Checks

We examine four alternative explanations for the grounding results: chance performance induced by candidate-set size, spatial proximity, dependence on image input, and sensitivity to candidate order. Rule-based baselines quantify performance attainable without model inference, whereas input controls compare each model before and after a targeted intervention. We therefore report them separately. Differences between model conditions follow the stated intervention protocol; model–

Table 13: Grounding accuracy of rule-based baselines (%).

Baseline	GAcc ↑	Scope and interpretation
Uniform random candidate	18.69	Chance level under the official aggregation weights
Nearest-region heuristic	**76.82**	Restricted to probes with a valid spatial distance

Note. Gray marks the chance reference; green bold highlights the strongest rule-based baseline.

baseline contrasts additionally match the observed probe support and aggregation weights. Spatial results are restricted to the rule's applicable pool.

Rule-Based Baselines. The original random-candidate expectation is 18.69%, and the nearest-region rule reaches 76.82% within its applicable subset. These summary scores use different eligibility pools. The matched comparisons below therefore recompute both baselines on exactly the model's evaluation support.

Same-support comparison and a geometry-defined hard subset. Let $\mathcal{H} \subseteq \mathcal{G}$ be the fixed three-source spatial pool on which the nearest-region rule applies (665 parents, 1,639 probes). This applicability set is distinct from the full grounding release and from the paired analysis sets. The nearest-region rule, candidate sets, and applicability decisions are frozen independently of model predictions. With h_j indicating whether the rule selects the correct candidate, define

$$\mathcal{H}_{\mathrm{hard}} = \{j \in \mathcal{H} : h_j = 0\}. \tag{13}$$

This subset (*H-hard*) contains 271 parents and 394 probes. Its membership is model-independent but uses reference targets to identify rule failures; it is not a rule-independent difficulty split. Nearest-region accuracy is zero on every subset of $\mathcal{H}_{\mathrm{hard}}$ by construction.

To distinguish applicability from response coverage, let $\mathcal{H}_m = \{j \in \mathcal{H} : A^{\mathrm{base}}_{m,j} = 1\}$ be model m's observed spatial support, using the returned-response convention of Appendix B.5. The observed hard and joint supports are

$$\mathcal{H}_{\mathrm{hard},m} = \mathcal{H}_{\mathrm{hard}} \cap \mathcal{H}_m, \qquad \mathcal{I}_m = \mathcal{L} \cap \mathcal{H}_{\mathrm{hard},m}.$$

Returned output failures are retained under the scoring contract; items without a response are not silently counted as observed. Tables 14 and 15 report the four open-weight models. For these rows, $\mathcal{H}_m = \mathcal{H}$ (665/1,639), $\mathcal{H}_{\mathrm{hard},m} = \mathcal{H}_{\mathrm{hard}}$ (271/394), and $\mathcal{I}_m = \mathcal{L} \cap \mathcal{H}_{\mathrm{hard}}$ (69/74), where each pair counts parents/probes. These are observed supports, not nominal pool sizes substituted for missing model outputs.

Each model–baseline comparison uses identical probe membership, candidate sets, and balanced weights (Appendix B.3). On the stated comparison set $\mathcal{S}$, random and nearest-region accuracies are $\mathrm{Agg}_{\mathcal{S}}(1/K_j)$ and $\mathrm{Agg}_{\mathcal{S}}(h_j)$, respectively, where K_j is the candidate count. With these inputs and weights fixed, the baseline scores are identical across models sharing $\mathcal{S}$; equal parent/probe counts alone do not establish identical support. The 19.84% random expectation on $\mathcal{H}$ and 18.76% on $\mathcal{H}_{\mathrm{hard}}$ therefore do not replace the 18.69% reference on the original random-eligible pool (Table 13). All intervals below use 2,000 source-parent-cluster bootstrap resamples (seed 2027), retaining records sharing a source parent together and recomputing the complete aggregation and same-support contrasts.

The nearest-region rule is competitive: it exceeds the smaller Qwen models and InternVL3.5-8B on identical support. Qwen3.5-27B nevertheless exceeds it by 9.53 points [6.10, 13.06]. On $\mathcal{H}_{\mathrm{hard}}$, this model obtains 73.22% grounding accuracy against 18.76% random accuracy, a 54.47-point margin [48.05, 60.45]. Thus, its grounding performance is not exhausted by this particular proximity rule. This comparison does not identify its internal strategy or exclude other spatial or non-spatial shortcuts.

Reading–grounding disagreement persists where the rule fails. Both the C^+G^- share and $\mathrm{GErr} \mid C$ in Table 15(b) use the same joint support $\mathcal{I}_m$ and weights:

$$q^{\mathrm{hard}}_{10,m} = \mathrm{Agg}_{\mathcal{I}_m}\left(C_m(1 - G_m)\right), \qquad \mathrm{GErr}^{\mathrm{hard}}_{m|C} = \frac{q^{\mathrm{hard}}_{10,m}}{\mathrm{Agg}_{\mathcal{I}_m}(C_m)}. \tag{14}$$

For Qwen3.5-27B, these quantities are 29.71% [20.14, 39.84] and 34.75% [23.77, 46.55], respectively. The first is a weighted event share over the intersection, whereas the second conditions on correct

Table 14: Same-support spatial comparison on $\mathcal{H}$ (665 parents/1,639 probes for every row). Accuracies are percentages; model-minus-nearest differences are percentage points with 95% intervals.

Model	Random $GAcc$ ↑	Nearest $GAcc$ ↑	Model $GAcc$ ↑	Model–Nearest (pp)
Qwen3.5-4B	19.84	76.82	49.04	-27.79 $[-32.16, -23.43]$
Qwen3.5-9B	19.84	76.82	51.78	-25.04 $[-29.12, -20.81]$
Qwen3.5-27B	19.84	76.82	**86.35**	**+9.53** $[6.10, 13.06]$
InternVL3.5-8B	19.84	76.82	30.30	-46.52 $[-50.58, -42.20]$

Bold highlights the model exceeding the tested rule on matched support. Values and differences are rounded independently. These subset scores are not overall $GAcc$ on $\mathcal{G}$.

Table 15: Diagnostics on the nearest-region failure subset. All shares and accuracies are percentages; differences are percentage points. Brackets give 95% source-parent-cluster bootstrap intervals.

(a) Grounding on $\mathcal{H}_{\text{hard}}$: 271 parents/394 probes

Model	Random $GAcc$ ↑	Model $GAcc$ ↑	Model–Random (pp)
Qwen3.5-4B	18.76	38.03	+19.27 $[12.38, 25.89]$
Qwen3.5-9B	18.76	35.21	+16.46 $[9.86, 23.09]$
Qwen3.5-27B	18.76	**73.22**	**+54.47** $[48.05, 60.45]$
InternVL3.5-8B	18.76	17.80	-0.96 $[-6.28, 4.56]$

(b) Reading–grounding on $\mathcal{L} \cap \mathcal{H}_{\text{hard}}$: 69 parents/74 probes

Model	C^+G^- share	GErr | C ↓
Qwen3.5-4B	27.54 $[17.42, 37.88]$	47.50 $[33.82, 62.00]$
Qwen3.5-9B	44.20 $[32.14, 56.72]$	62.24 $[49.44, 75.27]$
Qwen3.5-27B	29.71 $[20.14, 39.84]$	34.75 $[23.77, 46.55]$
InternVL3.5-8B	2.17 $[0.00, 5.88]$	100.00 —

Nearest-region $GAcc$ is zero on $\mathcal{H}_{\text{hard}}$ by definition. Panel (b) uses only its joint-reading intersection, not all 394 probes. The dash denotes an unreported interval because the conditional statistic is undefined in some resamples; it does not denote zero uncertainty. Panel (a) bold highlights the strongest margin over random selection. Conditional rates are diagnostic, not standalone cross-model rankings.

reading. InternVL3.5-8B's small event share does not imply reliable grounding: its conditional point estimate is 100%, and the reading mass vanishes in some resamples. The analysis establishes that the observed recognition–grounding discrepancy persists on this rule's failure subset, without claiming that all geometric shortcuts have been excluded.

No-Image and Candidate-Order Controls. The no-image control sends no image attachment while retaining the original problem text, evidence coordinates, and candidate geometry. It tests whether grounding performance depends on image pixels given the remaining non-pixel cues. The candidate-order control permutes only the candidate array; candidate aliases remain bound to the same physical regions. This intervention tests whether a fixed presentation order is sufficient to explain aggregate performance.

Table 16: Grounding under input controls (%); Δ denotes percentage-point change from the reference condition.

Model	Reference ↑	No image ↑	Δ_{img} (pp)	Reordered ↑	Δ_{ord} (pp)
Qwen3.5-4B	40.21	38.78	−1.43	41.93	+1.72
Qwen3.5-9B	40.52	43.53	+3.01	43.02	+2.50
Qwen3.5-27B	84.11	63.72	−20.39	84.48	+0.37
InternVL3.5-8B-HF	23.52	49.87	+26.35	7.24	−16.28
GPT-6 Astra	81.94	75.57	−6.37	84.27	+2.33
Gemini-3.8-flash	**90.71**	**76.90**	−13.81	**90.82**	+0.11

Note. Green bold / yellow shading indicate best / second-best accuracy within each accuracy column. Light red highlights large intervention responses ($|\Delta| > 10$ pp); Δ columns are diagnostic and therefore have no monotonic better direction.

Removing the image separates the models by reference performance. The three models with reference GAcc above 80% all decline: Qwen3.5-27B (Team, 2026), GPT-6 Astra (OpenAI, 2026b), and Gemini-3.8-flash (Google DeepMind, 2026) lose 20.39, 6.37, and 13.81 percentage points, respectively. Their declines indicate sensitivity to image evidence beyond the retained text and candidate metadata, but do not rule out spatial shortcuts. The weaker models behave differently. Qwen3.5-4B (Team, 2026) changes by only -1.43 points, while Qwen3.5-9B (Team, 2026) and InternVL3.5-8B-HF (Wang et al., 2025b) improve by 3.01 and 26.35 points. These divergent responses distinguish access to visual evidence from the ability to use it: attaching an image does not ensure that the visual interface contributes useful information. Because the control retains text, coordinates, and candidate geometry, the differences measure sensitivity to image input rather than the isolated causal contribution of visual reasoning.

Candidate reordering rules out a fixed ordering effect for five of the six models. Their GAcc changes by at most 2.50 points, indicating that candidate position does not drive their aggregate grounding performance. InternVL3.5-8B-HF (Wang et al., 2025b) is the exception, falling from 23.52% to 7.24% (-16.28 points). Together with its large improvement without the image, this decline indicates pronounced sensitivity to input presentation and warrants separate interpretation of its grounding score. Similar aggregate accuracy does not imply identical item-level predictions, but the control is sufficient to reject fixed candidate order as a complete explanation for the other five models.

Across the four controls, neither random selection nor a general candidate-order bias is sufficient to explain the PhysAlign results. The strongest models use image evidence, while spatial proximity remains a strong but incomplete baseline. The anomalous response of InternVL3.5-8B-HF (Wang et al., 2025b) also shows that interface robustness is a model property that should be measured rather than assumed. These experiments narrow the explanations based on candidate structure, geometry, and input formatting; they do not identify a model's internal strategy or exclude shortcuts outside the present controls.

D.3 Grounding and Original-Problem Solving

This section documents the grouping behind Section 5.4 and Figure 5. The matched analysis contains 311 parents and 412 local probes from the full release; it is not the denominator of $SolveAcc$, which uses all 986 parents. Matching links original-problem outcomes to local-probe records and is distinct from the eligibility definition of $\mathcal{L}$, even though their reported counts coincide. Solving and probing use independent contexts, without shared history, predictions, or scoring feedback.

The solve-correct and solve-wrong groups refer to the original binary final-answer correctness outcomes. These labels are distinct from the normalized partial-credit scores s_i in Appendix B.4; $SolveAcc$ is not the proportion of parents in the solve-correct group. For each model, local grounding is averaged within each parent and then across parents in each binary-outcome group, yielding $\bar{G}_{\text{wrong}}$ and $\bar{G}_{\text{correct}}$. We report $\Delta_{\text{solve}} = \bar{G}_{\text{correct}} - \bar{G}_{\text{wrong}}$. Table 17 retains these outcome-conditioned estimates.

These outcome-conditioned means use parent-level averaging and need not equal the type-balanced $GAcc$ in Table 2. Group membership is determined separately for each model and may reflect problem difficulty and other unobserved factors; the solve-correct and solve-wrong subsets are not fixed across models. The reported differences are descriptive point estimates, which alone establish neither statistical significance nor causality. Independent probing also cannot establish which local

Table 17: Parent-level grounding by binary final-answer correctness (311 parents, 412 probes; %), distinct from partial-credit $SolveAcc$. Δ_{solve} uses unrounded group means and is reported in percentage points.

Model	$\bar{G}_{\text{wrong}}$	$\bar{G}_{\text{correct}}$	Δ_{solve} (pp)
Qwen3.5-4B	39.19	41.31	+2.12
Qwen3.5-9B	40.67	44.61	+3.94
Qwen3.5-27B	85.51	87.89	+2.39
InternVL3.5-8B	23.37	25.88	+2.51
GPT-6 Astra	83.05	85.53	+2.48
Gemini-3.8-flash	92.08	92.36	+0.28

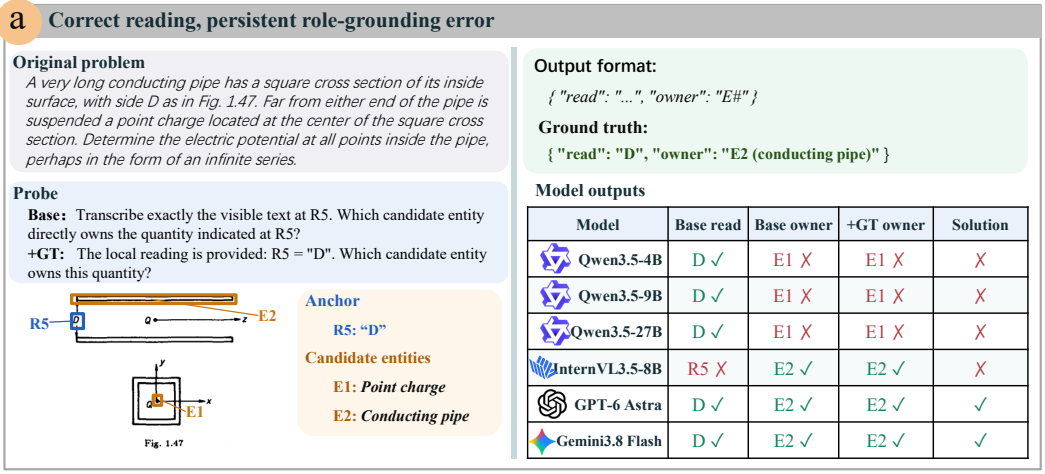

Figure 9: Case 1: A role mismatch persists despite correct reading. All three Qwen3.5 models correctly read D at anchor R5, yet assign it to the point charge E1 rather than the conducting pipe E2 under both Base and +GT.

evidence was used during solving or whether correcting a mismatch would alter the final answer. The cases in Appendix D.4 illustrate these observable distinctions without estimating their prevalence.

D.4 Representative Case Studies

We use four complementary cases to characterize the relationship among content recognition, physical-role grounding, and original-problem solving. Each case reports the local evidence, candidate entities, predictions under the Base and +GT conditions, and the outcome of an independent run on the original problem. The first three cases isolate distinct diagnostic patterns; the fourth connects an observed grounding error to its equation-level consequence through a controlled analytical construction. Binary solution labels in these figures indicate final-answer outcomes, not the amount of partial credit. These cases are diagnostic rather than estimates of error frequency: they distinguish behaviors that may yield the same solution-level outcome but reflect different failure modes.

Figure 9 directly instantiates the $C^{+}G^{-}$ pattern discussed in the main text. Qwen3.5-4B (Team, 2026), Qwen3.5-9B (Team, 2026), and Qwen3.5-27B (Team, 2026) all transcribe the side-length symbol D correctly but select the wrong physical owner, E1. Their predictions remain unchanged when +GT explicitly supplies the correct reading. The failure therefore lies beyond local transcription, in the association between the recognized content and its physical entity. GPT-6 Astra (OpenAI, 2026b) and Gemini-3.8-flash (Google DeepMind, 2026) recover both the reading and the owner from the same evidence, providing evidence that the intended relation is recoverable from the input. Correct recognition alone does not guarantee correct physical-role grounding.

Figure 10 shows a complementary paired outcome. Anchor R1 marks the vertical-axis label $v/(\mathrm{m/s})$ of object B's v–t graph, with owner E2. Under Base, InternVL3.5-8B (Wang et al., 2025b) reports x/m and selects E1; under +GT, its owner prediction changes to E2. This transition is consistent with a reading-linked error, but does not establish causal mediation by recognition. Together with Figure 9, it illustrates that supplied readings can accompany either repaired or persistent role errors.

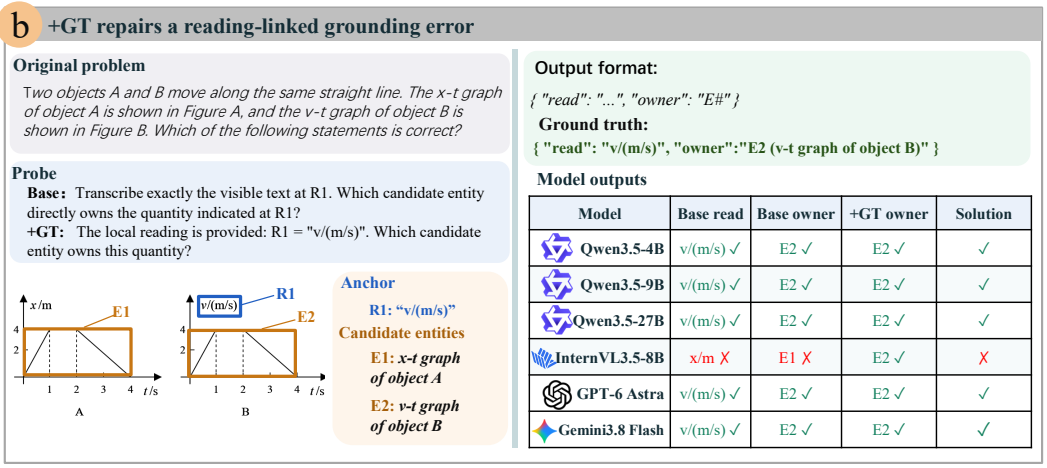

Figure 10: Case 2: A grounding error is repaired under +GT. Under Base, InternVL3.5-8B (Wang et al., 2025b) misreads the vertical-axis label and selects E1; after receiving the correct reading, it changes its owner prediction to E2.

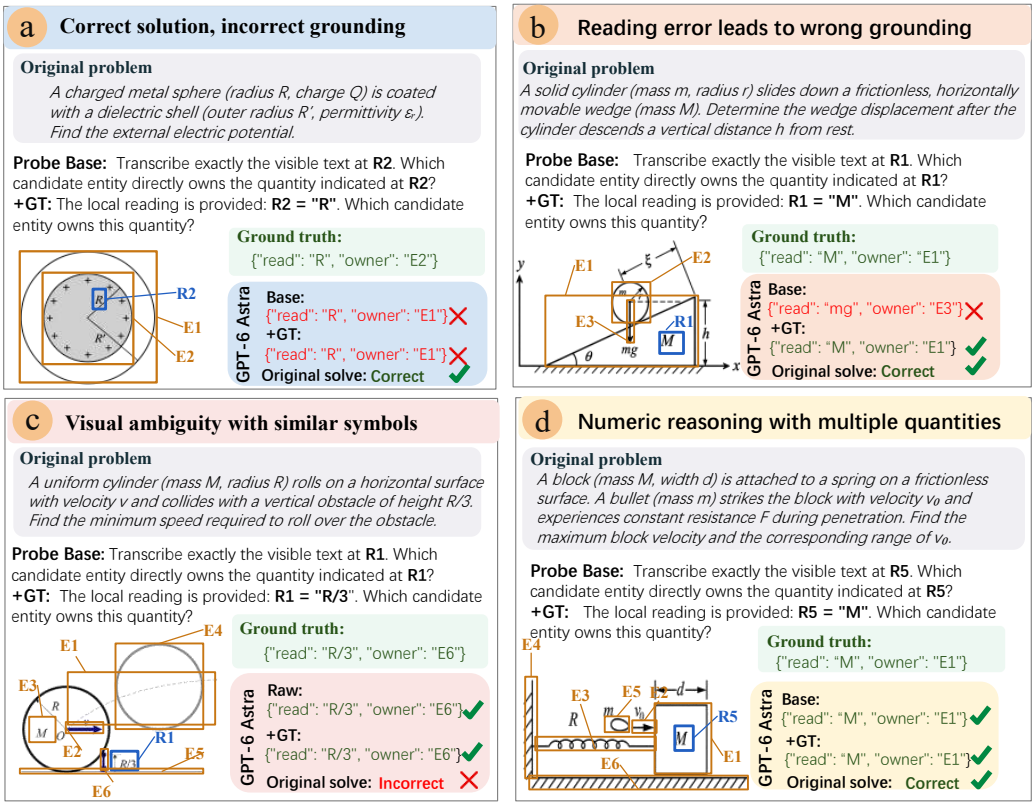

Figure 11: Case 3: Local grounding and original-problem solving are not equivalent measures. Panels (a) and (c) show a correct solution with incorrect grounding and correct grounding with an incorrect solution, respectively; panel (d) shows a consistent success.

The independent solving response is incorrect; because that run did not receive the +GT input, this case does not test whether the added reading would improve problem solving.

Figure 11 establishes the bidirectional non-equivalence between the local diagnostic and original-problem solving. In panel (a), GPT-6 Astra (OpenAI, 2026b) solves the original problem correctly but consistently assigns R' to the wrong entity E1 under both Base and +GT. A correct final answer

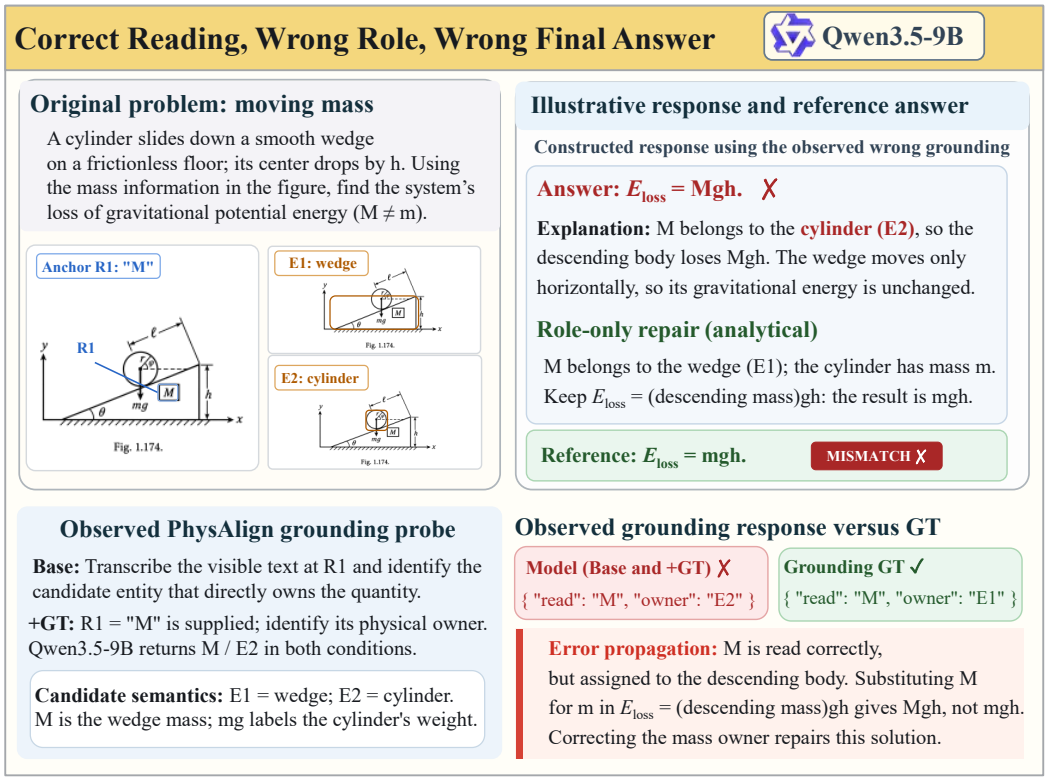

Figure 12: Case 4: A recorded role mismatch has a concrete equation-level consequence under controlled substitution. Qwen3.5-9B (Team, 2026) reads M correctly but assigns it to the descending cylinder E2 rather than the wedge E1 under both Base and +GT.

thus does not imply consistent recovery of local physical roles. Panel (c) exhibits the converse: the model correctly reads and grounds $R/3$, yet answers the original problem incorrectly, showing that one correct grounding decision cannot rule out subsequent modeling or derivation errors. Panel (b) provides another instance in which +GT repairs a reading-mediated error, whereas panel (d) shows correct recognition, grounding, and solving in a scene containing several annotated quantities. The probe therefore distinguishes consistent success from reading errors, role mismatches, and downstream reasoning failures rather than merely introducing additional opportunities for failure.

Figure 12 links the observed probe output to a specific physical consequence. The recorded grounding response from Qwen3.5-9B (Team, 2026) is $\{\text{read} : M,\ \text{owner} : \text{E2}\}$ in both conditions, although M denotes the mass of the wedge E1 and the descending cylinder has mass m. The analytical construction on the right keeps the valid relation $E_{\text{loss}} = (\text{descending mass})gh$ fixed and changes only the quantity selected by the reported ownership. Substituting M yields Mgh, whereas correcting the owner restores the reference result mgh. A role mismatch can therefore alter which physical quantity enters an otherwise valid derivation. The constructed response is not treated as the model's observed solution trace; it is a controlled counterfactual that isolates the consequence of the observed grounding output.

Across the four cases, content recognition, physical-role grounding, and original-problem solving are related but distinct observable outcomes: a recognition error may induce a role mismatch, a correct reading may coexist with a persistent ownership error, and a correct final answer may conceal a local premise-level inconsistency. Under the controlled substitution in Case 4, an ownership error also changes the variable used in an otherwise unchanged equation. PhysAlign therefore complements answer-level evaluation with localized diagnostics of these intermediate relations. Because the local probes and original solutions are executed in separate contexts, the results establish behavioral non-equivalence rather than a causal account of the model's internal reasoning. The controlled construction demonstrates what follows if the reported ownership is substituted into a fixed derivation; it does not show that the model used this premise in its independently generated solution.

E Comparison with Related Benchmarks

We extend Section 2 by distinguishing problem-level reasoning, structured visual understanding, and controlled error diagnosis.

Problem- and step-level reasoning. MathVista (Lu et al., 2024) evaluates mathematical reasoning in visual contexts, and PhysUniBench (Wang et al., 2025a) focuses on undergraduate physics. PhysReason (Zhang et al., 2025) evaluates both final answers and intermediate solution steps. PhysAlign instead tests local content–role correspondences, without requiring complete solutions or treating solution-derived facts as probe targets.

Diagram structure, relations, and reading. AI2D (Hiippala et al., 2021) represents diagram constituents and relations as diagram parse graphs. GQA (Hudson & Manning, 2019) constructs compositional questions from scene graphs and measures consistency and grounding. TextVQA (Singh et al., 2019) requires reading and reasoning about image text. PhysAlign likewise emphasizes structured visual evidence, but provides fixed anchors and candidates and scores literal recognition separately from physical-role correspondence. It neither requires complete diagram parsing nor equates role grounding with visual localization.

Controlled error diagnosis. HallusionBench (Guan et al., 2024) uses controlled image–question groups to diagnose language hallucination and visual illusion. PhysAlign tests whether supplying a verified local reading changes the reported role, keeping the original problem, anchor, candidates, and scoring fixed. This control neither removes all perceptual uncertainty nor reveals internal reasoning.

Table 18 compares the protocols of selected closely related benchmarks. Its criteria concern diagnostic interfaces, not overall benchmark quality. Associations between PhysAlign scores and independently evaluated problem solving are descriptive, not causal.

Table 18: Protocol-level comparison with closely related benchmarks.

Benchmark	General evaluation		Localized diagnosis				Utility
	Input Ctrl.	Interm. Eval.	Fixed Anchor	Rec./Grd. Split	Paired Rec. Ctrl.	Det. Local Score	Solve Link
SeePhys (Xiang et al., 2026a)	✓	✗	✗	✗	✗	✗	✗
SeePhys Pro (Xiang et al., 2026b)	✓	△	✗	✗	✗	✗	✗
PhysicsArena (Dai et al., 2025)	✗	✓	✗	✗	✗	✗	✓
MathVerse (Zhang et al., 2024)	✓	✓	✗	✗	✗	✗	✗
PhysAlign (Ours)	✓	✓	✓	✓	✓	✓	✓

Legend: ✓ explicit; △ indirect or related; ✗ no corresponding protocol under the definitions below. Symbols indicate protocol features, not model performance or overall benchmark quality. **Input Ctrl.**: controlled changes to evidence availability or modality. **Interm. Eval.**: evaluation of intermediate reasoning stages or outputs. **Fixed Anchor**: evaluation at a predefined local image region or text span in an unchanged source problem. **Rec./Grd. Split**: separate recognition and role-grounding scores on the same probe. **Paired Rec. Ctrl.**: matched probes with and without correct local readings, with the grounding target withheld. **Det. Local Score**: rule-based local scoring without an LLM judge. **Solve Link**: association between separately evaluated diagnostic scores and original-problem correctness. For SeePhys Pro, △ denotes modality-transfer-based diagnosis rather than direct intermediate-output scoring.

F Scope, Limitations, and Release Documentation

Measurement Scope. PhysAlign measures whether a model can map locally visible content to its correct physical role when given evidence anchors, candidate entities, and a constrained output interface. The protocol scores content recognition and role grounding separately, but it does not assess unconstrained, holistic diagram understanding. The +GT condition supplies only a verified local transcription: it reveals neither the correct candidate nor the role label, and it leaves uncertainty in entity boundaries, spatial relations, and physical semantics unresolved. The Base/+GT comparison

therefore measures the change in grounding after a correct transcription is supplied; it is not an upper bound under perfect perception.

Interpretive Boundaries. Our measurements concern explicit local correspondences reported under fixed inputs. They do not recover a model's latent representation, complete physical world model, or internal reasoning process. Local probes and original-problem solving are evaluated in separate contexts. Agreement or disagreement between them supports a distinction at the level of observable behavior, but cannot determine which local premise the model used, ignored, or propagated incorrectly in a particular solution. Likewise, a controlled substitution based on an incorrect role characterizes a possible consequence within a fixed derivation; it does not causally identify the model's actual reasoning trace.

Coverage and Release Boundaries. The current evaluation focuses on static educational physics diagrams and is skewed toward mechanics and electromagnetism. Retained probes cover reviewed local facts rather than every expressible relation in each problem, and paired +GT inputs are available only for a subset. Because the parent problems come from public datasets, we cannot exclude prior exposure to the original problems or semantically similar variants during pretraining. Parent-level split propagation keeps probes from one parent together, and exact-input deduplication removes identical public inputs; neither certifies a contamination-free evaluation. The reported experiments combine development and test within the retained release and should therefore be interpreted as release-derived diagnostics, not held-out test estimates. Appendix A documents provenance, review, retention, and coverage. These documentation priorities are consistent with Datasheets for Datasets (Gebru et al., 2021), which advocates explicit reporting of dataset composition, collection, and recommended uses. We plan to release the probe inputs and metadata, prompt templates, evaluation code, and recorded run configurations for which redistribution is permitted, while keeping public inputs separate from private reference targets and scoring mappings. Source problems and images will be redistributed only where upstream licenses permit; otherwise, the release will provide provenance identifiers and reconstruction instructions without asserting additional rights over third-party content.