One Relation Is Not Enough: Benchmarking Relational Spatial Understanding in Vision-Language Models

Anonymous authors
Paper under double-blind review

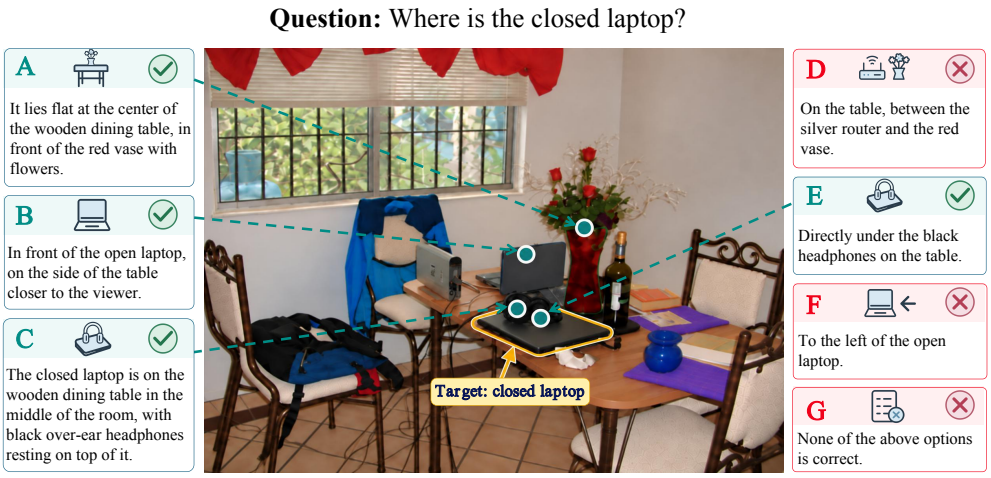

Figure 1: A representative multi-select question assessing coexisting spatial relations. The closed laptop admits four valid descriptions grounded in different anchors and levels of detail (A, B, C, and E), while D and F contradict the image. Because valid ordinary answers exist, the "None of the above options is correct" option G is also incorrect. The task requires predicting the complete set $\{A, B, C, E\}$ rather than finding any single plausible answer.

Abstract

Spatial understanding requires more than recognizing isolated spatial relations; it also requires understanding how diverse descriptions involving different reference objects and reference frames jointly characterize the same spatial structure. A correct answer under a single formulation is therefore insufficient to reflect a model's ability to identify multiple valid spatial descriptions completely and accurately. We introduce CoSpatial-Bench, a direct multi-select benchmark that evaluates whether vision-language models can select all valid descriptions of the same target from a fixed candidate set, based on the scene and question, while rejecting unsupported options. The Base variant contains 495 questions and 2,000 human-reviewed ordinary candidate descriptions. Our primary metric, *Error-Free Coverage (EFC)*, measures coverage of valid descriptions under a no-false-selection constraint, while paired variants examine recognition performance under changes in referring expressions, viewpoints, and spatial conditions. Across ten VLM baselines, the best Base EM is only 33.54, while the highest EFC is 42.53. A separate single-candidate intervention study characterizes correction of initially incorrect option decisions under perceptual support and subsequent combined perceptual and reasoning support: across three diagnostic cohorts, perceptual support corrects 58.7–70.7% of initial errors,

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.

while 8.6–13.8% remain unresolved after both stages. We further introduce *Target-Centered Relational QA Synthesis* (TCR-QAS) and train Qwen3.5-4B with whole-question GRPO on the synthesized data, increasing its Base EFC from 28.17 to 33.37. These results highlight the need to evaluate spatial understanding beyond the correctness of a single answer: complete and error-free identification of diverse valid descriptions remains challenging.

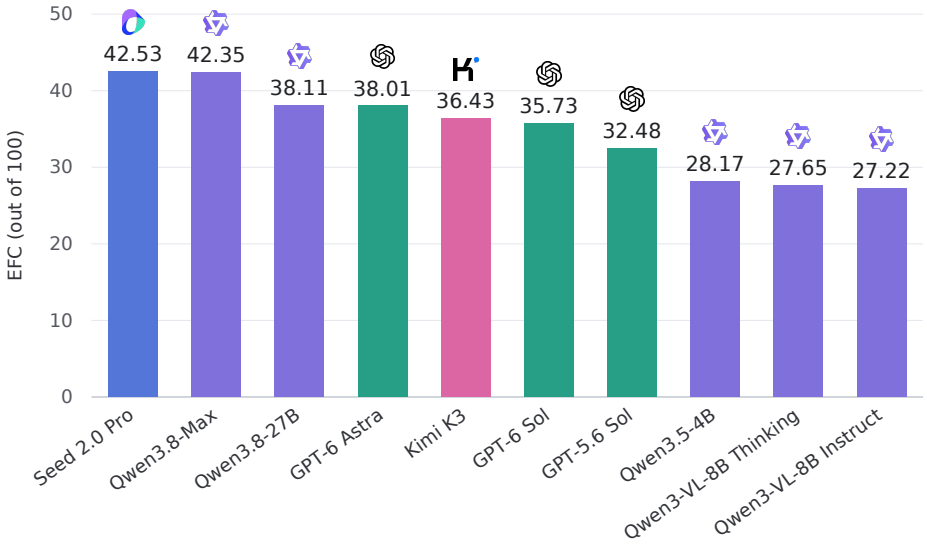

Figure 2: Base benchmark results on 495 questions. Models are ranked by EFC (higher is better). Doubao Seed 2.0 Pro achieves the highest EFC of 42.53, followed by Qwen3.8-Max (42.35) and Qwen3.8-27B (38.11). Full results are reported in Table 2.

1 Introduction

Spatial understanding enables vision-language models (VLMs) to locate objects, interpret scenes, and reason across viewpoints. For a question about one target, several spatial descriptions can be simultaneously valid, differing in relation, reference object (anchor), and level of specificity. Correctly recognizing one description does not establish whether a model can identify the others while rejecting unsupported alternatives.

Existing benchmarks assess relation validity, spatial question answering, and situated or compositional reasoning (Yang et al., 2019; Ma et al., 2023; Gong et al., 2026). Some require integrating multiple relations or selecting multiple answers. These requirements do not by themselves assess complete and sound recognition of coexisting descriptions for the same target and question.

Figure 1 illustrates the distinction. The closed laptop is in front of the red vase, in front of the open laptop, and below the headphones. These descriptions differ in anchor and relation while locating the same target under the same reference frame; they are not merely paraphrases. Selecting one correct description demonstrates local correctness without establishing recognition of the complete valid set. In this sense, *one relation is not enough*: evaluation requires both *coverage* of valid supplied descriptions and *soundness* in rejecting unsupported ones.

We introduce CoSpatial-Bench, a direct multi-select benchmark that evaluates these requirements by asking models to jointly select all valid descriptions from a fixed candidate set while rejecting unsupported options. Its Base variant contains 495 questions and 2,000

human-reviewed candidate descriptions. Our primary metric, *Error-Free Coverage (EFC)*, measures valid-description coverage under a no-false-selection constraint.

Across ten VLM baselines, the highest EM is only **33.54**, achieved by Qwen3.8-Max; the highest EFC is **42.53**, achieved by Doubao Seed 2.0 Pro (Figure 2). This result shows that complete valid-set identification remains difficult for the evaluated models.

We assess sensitivity to referring expressions, viewpoints, and spatial conditions through paired Base and Augmented questions. We also vary option order to check whether results depend on answer positions. A separate option-level intervention study characterizes correction of initial errors under staged perceptual and reasoning support. Finally, we use *Target-Centered Relational QA Synthesis* (TCR-QAS) to construct a separate 1,200-question training subset and evaluate EFC-aligned GRPO training of Qwen3.5-4B (Qwen Team, 2026a; Shao et al., 2024). Its Base EFC increases from 28.17 to 33.37, with gains across both benchmark variants and option orders.

Our contributions are threefold: (1) a benchmark and direct multi-select formulation for jointly identifying coexisting valid spatial descriptions through coverage and soundness; (2) a diagnostic characterization of incorrect option decisions under staged perceptual and reasoning support; and (3) TCR-QAS and a GRPO study showing improved EFC for Qwen3.5-4B under the benchmark protocol.

2 Related Work

Spatial reasoning benchmarks. Spatial evaluation tests whether models ground spatial language in scenes, from relation validation to question answering in 3D environments (Yang et al., 2019; Liu et al., 2023; Azuma et al., 2022). Compositional and situated benchmarks extend this goal to combining relations and interpreting specified viewpoints (Hudson & Manning, 2019; Ma et al., 2023; Liu et al., 2025; Li et al., 2026). Together, these directions cover varied spatial scopes, viewpoints, and reasoning demands. Their emphasis is on the queried relation or requested answer; integrating several relations to produce that answer does not itself require selecting other valid descriptions of the same target.

Multiple valid spatial descriptions. Closer precedents evaluate multiplicity through relation labels or compositional answer choices. SpatialVOC2K annotates multiple applicable prepositions for an ordered object pair, but its reported top-n evaluation credits a hit among reference labels rather than complete label-set selection (Belz et al., 2018). SpaCE-10 includes compositional questions with five options and two correct answers, requiring both to be selected (Gong et al., 2026). Thus, multi-label annotations and complete multi-answer scoring have relevant precedents; the distinction for CoSpatial-Bench is the description space evaluated. Supplied candidates answer the same question about the same target but may differ in relation, anchor, and specificity, and each may contain multiple claims under the applicable reference frame. The number of valid options varies, and evaluation jointly considers coverage and soundness. SpatiaLQA instead evaluates generated procedural steps and their preconditions (Xie et al., 2026), a different output requirement from identifying supplied spatial descriptions. Appendix A provides the detailed comparison.

Diagnostic interventions and spatial learning. Evidence-supported diagnosis examines how spatial judgments respond to bounding boxes, scene graphs, and structural inputs (Shiri et al., 2024; Li & Yu, 2026). Our option-level study characterizes correction of initially incorrect option decisions from multi-select responses under staged perceptual and reasoning support, without treating correction as a causal decomposition of error sources. Learning approaches seek to improve spatial performance through dedicated supervision (Chen et al., 2024; Ray et al., 2025) or spatial reinforcement learning (Ma et al., 2025b). Our TCR-QAS and GRPO study specifically tests whether training with an EFC-aligned reward improves joint valid-set selection under the benchmark protocol.

3 Benchmark Construction and Evaluation Protocols

3.1 Task Formulation

We formulate relational spatial understanding in CoSpatial-Bench directly as a multi-select question. Each instance presents an image I, a question q, and the complete set of ordinary candidate descriptions $A = \{a_1, \ldots, a_m\}$. Let $G_{\text{ord}} \subseteq \{1, \ldots, m\}$ denote the indices of all correct ordinary candidates. Every question also includes an option $\perp$ labeled "None of the above options is correct"; the gold answer set is

$$G = \begin{cases} G_{\text{ord}}, & G_{\text{ord}} \neq \varnothing, \\ \{\perp\}, & G_{\text{ord}} = \varnothing. \end{cases} \tag{1}$$

The model receives the image, question, and all options together and selects an answer set $S \subseteq \{1, \ldots, m, \perp\}$ in a single response. Scoring uses the selected set directly.

Option $\perp$ is correct if and only if no ordinary candidate is correct. Selecting $\perp$ together with any ordinary option is invalid and receives zero credit. An empty prediction is distinct from selecting $\{\perp\}$: it receives zero EM and EFC, with BA defined in Section 4.3.

Candidate validity. A candidate is considered correct when its material claims are supported by the image under the applicable reference frame and it answers the question. For localization questions, a description must meaningfully constrain the target object's location, without necessarily identifying a unique position. Different candidates may therefore provide simultaneously valid descriptions through different reference objects, spatial relations, or levels of detail. A candidate can contain multiple spatial claims; the complete description is the unit of labeling and scoring. For gold labeling, each candidate is checked against its visual evidence and the viewpoint or coordinate system specified by the question. The model then selects among all options jointly.

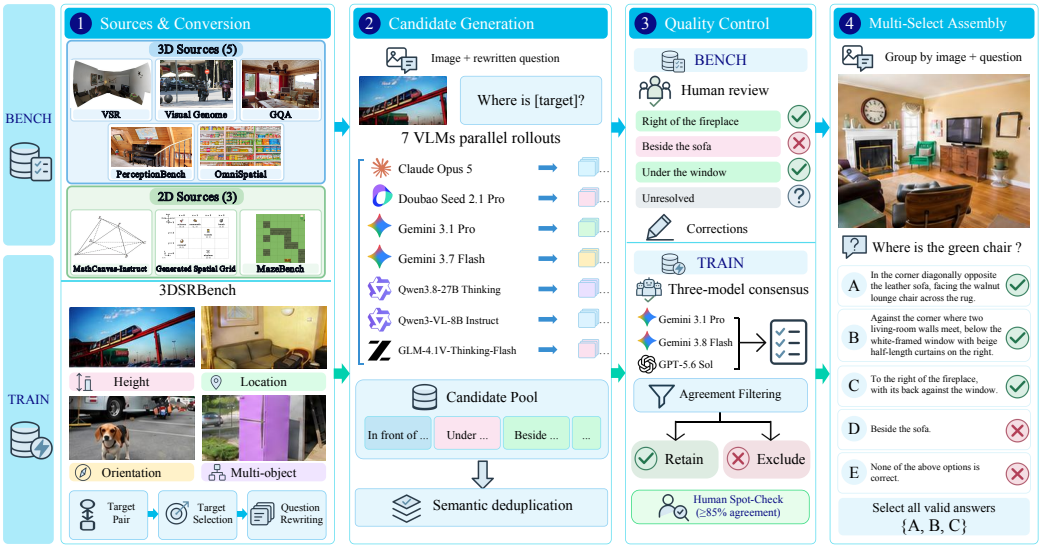

Figure 3: Construction pipelines for CoSpatial-Bench and the RL training data, from candidate generation and quality control to assembly into multi-select questions.

3.2 Data Construction and Quality Control

Benchmark construction. Human annotators select images from eight sources, including synthetic two-dimensional grids, and determine questions about targets that admit multiple valid spatial descriptions. For each image and question, seven VLMs generate candidate descriptions through rollouts: GLM-4.1V-Thinking-Flash, Claude Opus 5, Doubao

Seed 2.1 Pro, Gemini 3.1 Pro, Gemini 3.7 Flash, Qwen3-VL-8B Instruct, and Qwen3.8-27B (Figure 3). We cluster candidates by semantic similarity; human reviewers then remove descriptions with identical or near-identical meanings, retaining those with different reference objects or reference frames. Two reviewers jointly check correctness and discard disputed candidates. All 2,000 ordinary candidates in the final Base set pass this review and are grouped by image, question, and target into multi-select instances.

RL training data. We use *Target-Centered Relational QA Synthesis* (TCR-QAS) to construct a separate training set from relation-level 3DSRBench annotations. For each relational pair, we select a target object with informative spatial anchors and reformulate the original relation into a target-centered localization question. The same seven models listed above generate candidate descriptions. Gemini 3.1 Pro, Gemini 3.8 Flash, and GPT-5.6 Sol independently label each candidate. We retain candidates only when all three agree on the correctness label. The retained candidates are assembled into multi-select questions and difficulty-filtered to obtain the final 1,200-question GRPO training subset. Human reviewers audit 500 sampled questions from this subset; 85% of their candidate labels agree with the retained labels.

Dataset composition. Table 1 reports candidate labels and question composition; Figure 4 shows the Base scene and source mix.

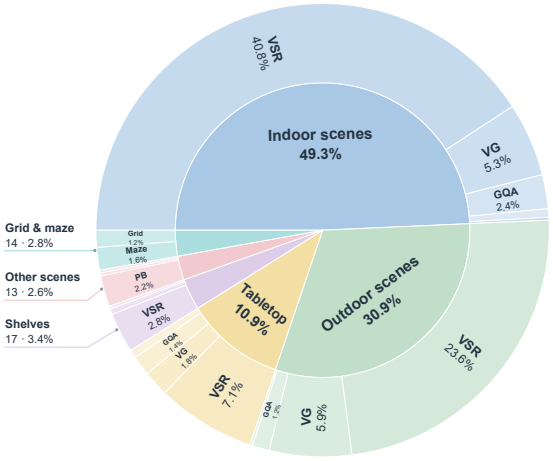

Figure 4: Base composition (495 questions): scene families (inner ring) and source datasets (outer ring). Indoor: 49.3%; outdoor: 30.9%; VSR: 369 questions (74.5%).

Table 1: Dataset sizes and label distributions.

Dataset	Size		Candidate labels (%)		Question composition (%)		
	Questions	Candidates	Correct	Incorrect	All correct	Mixed	All incorrect
Base	495	2,000	51.6	48.4	19.0	71.3	9.7
Augmented	495	2,021	48.8	51.2	17.6	70.9	11.5
RL training	1,200	5,136	52.8	47.2	9.1	85.4	5.5

3.3 Paired Transformations

We pair each Base question with an Augmented counterpart using one of three transformations: attribute-based reference, person-centered viewpoint, or a counterfactual two-dimensional mirror. Each version is scored against its own reviewed answer set. Appendix C details the transformation rules, pair alignment, and examples.

4 Experiments

Our experiments address two complementary questions: whether models recognize the multiple valid descriptions in a complete spatial question, and how incorrect ordinary-option decisions in those responses change under perceptual and reasoning information. The multi-select response provides the initial decisions; staged Ref-assisted single-candidate prompts support diagnostic attribution. We further evaluate whether GRPO training improves Qwen3.5-4B's valid-set identification. Training details and the respective prompt templates appear in Appendices F and G.

4.1 Baseline Models

We evaluate ten VLM baselines on Base and Augmented and include a GRPO-trained Qwen3.5-4B checkpoint. Paired Augmented questions test sensitivity to question formulation and spatial conditions, while a second option ordering checks dependence on answer positions. Model settings and option-order details appear in Appendices G.1 and C.2.

4.2 Model Training

We train Qwen3.5-4B with GRPO (Shao et al., 2024) on a separate subset of 1,200 questions. For each question, we sample eight responses and score each with

$$R(S, G) = s(S, G)^2, \tag{2}$$

where s is the per-question EFC score before scaling to 0–100, as defined below. Squaring preserves the zero reward for an empty selection or any selected incorrect option, while reducing the reward for a correct but incomplete set. A complete valid-set prediction retains reward one; among error-free predictions, the squared objective therefore places greater weight on identifying the full set than the unsquared EFC score does. Because s divides by the number of valid options, the reward applies across questions with different gold answer set sizes. The squared score is used only for training; EM, BA, FSR, and EFC retain their definitions at evaluation.

Training uses a learning rate of 10^{-6}, batch size 32, and KL-loss coefficient 0.01, with a separate 95-question validation split. We compare baseline and post-training performance under the same scoring rules. Table 4 reports the main Base comparison; Appendix F.2 gives the remaining training settings, and Appendix F.3 reports EFC across variants and option orders.

4.3 Evaluation

Base and Augmented each contain 495 multi-select questions, scored against their respective reviewed answer sets. For each question, the model receives an image, a question, and all options, including the "None of the above options is correct" option, and selects those it judges correct. Let G_i and S_i denote the gold and selected sets for question i, respectively; G_i is always nonempty. EM, BA, and EFC include $\perp$; FSR excludes it. All metrics are reported separately for Base and Augmented.

Error-Free Coverage. We use Error-Free Coverage (EFC) as the primary metric:

$$s(S, G) = \begin{cases} \dfrac{|S|}{|G|}, & \varnothing \neq S \subseteq G, \\ 0, & \text{otherwise}, \end{cases}$$
$$\text{EFC} = \frac{100}{N} \sum_{i=1}^{N} s(S_i, G_i). \tag{3}$$

where $N = 495$ for both Base and Augmented. A prediction receives credit equal to the fraction of valid options selected, provided that every selected option is correct. Selecting

any incorrect option or returning an empty set yields zero. EFC weights questions equally and is reported on a 0–100 scale. For questions with no correct ordinary option, selecting $\perp$ alone receives full credit.

Exact match and balanced accuracy. We report exact match (EM) and balanced accuracy (BA) as complementary metrics. EM is the percentage of questions for which the predicted set exactly matches the gold answer set:

$$\mathrm{EM} = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}[S_i = G_i]\,. \tag{4}$$

BA measures recognition of correct options and rejection of incorrect ones. Let $\mathcal{O}_i$ denote all options for question i and $B_i = \mathcal{O}_i \setminus G_i$ its incorrect options. Including $\perp$ ensures that both G_i and B_i are nonempty. For a valid prediction, the true positive rate (TPR) is the fraction of correct options selected, and the true negative rate (TNR) is the fraction of incorrect options rejected:

$$\mathrm{TPR}_i = \frac{|S_i \cap G_i|}{|G_i|}, \qquad \mathrm{TNR}_i = \frac{|B_i \setminus S_i|}{|B_i|}. \tag{5}$$

We average these two rates within each question and then across questions:

$$b_i = \begin{cases} \dfrac{\mathrm{TPR}_i + \mathrm{TNR}_i}{2}, & \text{valid output,} \\ 0, & \text{invalid or failed output,} \end{cases}$$
$$\mathrm{BA} = \frac{100}{N} \sum_{i=1}^{N} b_i. \tag{6}$$

BA thus gives equal weight to recognizing correct options and rejecting incorrect ones. It captures option-level discrimination, while EM requires an exact answer set and EFC credits coverage only when the selected set contains no incorrect option.

False Selection Rate. We report option-level False Selection Rate (FSR) to measure acceptance of incorrect spatial descriptions. Let B_i^{ord} be the set of incorrect ordinary options for question i, and let $a_{ij} = 1$ if option j is accepted as correct and 0 otherwise. Then

$$\mathrm{FSR} = 100\, \frac{\sum_{i=1}^{N} \sum_{j \in B_i^{\mathrm{ord}}} a_{ij}}{\sum_{i=1}^{N} |B_i^{\mathrm{ord}}|}. \tag{7}$$

FSR pools option counts across questions and excludes $\perp$. The denominator is 969 for Base and 1,034 for Augmented. Each option is counted independently: an unresolved judgment does not erase a recorded decision on another option in the same question. Empty or failed responses remain in the denominator but contribute no false selections. Lower is better. FSR measures false acceptance, not correct rejection or answer coverage, and is interpreted alongside EM, BA, and EFC.

Staged evidence interventions. We derive the initial condition C_0 from each model's whole-question multi-select response on Base. For an ordinary candidate, selecting an incorrect option or omitting a correct option is an incorrect initial decision. For cases with both types of Ref information available, we first supply a Perception Ref in a single-candidate prompt (C_P). Candidates still judged incorrectly then receive both the same Perception Ref and a Reasoning Ref (C_{PR}). Each evaluation uses a fresh context with the same image, question, candidate, and selection criterion. GPT-6 Astra with high reasoning effort generates both Refs with access to the human correctness label. Human review is triggered only when the generator's conclusion disagrees with that label. The Refs supplied to the evaluated model contain neither the correctness label nor the gold answer set.

In the Ref-assisted prompts, `correct` means that the ordinary option should be selected: the candidate must answer the question and all key descriptions must be supported. A key

contradiction or an unsupported key claim prevents selection. We compare each decision with the same human-reviewed Base label used for the initial option decision.

We use the first agreement with that label to classify initial errors: *perceptual-support recovery* under C_P; *additional reasoning-support recovery* under C_{PR} for cases still incorrect under C_P; or *unresolved after both stages* if neither condition yields a correct decision. This operational attribution identifies the support condition associated with correction. It does not establish the original error source, since C_0 comes from a multi-select response whereas the follow-ups use single-candidate prompts. Appendix E.2 defines the denominator used for each rate.

5 Results

5.1 Performance on Base and Augmented Questions

Table 2: VLM performance on Base and Augmented.

Model	Base (%)				Augmented (%)			
	EM↑	BA↑	FSR↓	EFC↑	EM↑	BA↑	FSR↓	EFC↑
Doubao Seed 2.0 Pro High	28.48	65.39	**16.20**	**42.53**	25.25	64.68	**16.25**	39.22
Qwen3.8-Max	**33.54**	74.57	41.80	42.35	**32.12**	**74.49**	38.01	**41.66**
Qwen3.8-27B	27.07	71.86	39.94	38.11	25.25	71.07	30.08	39.37
GPT-6 Astra High	30.91	**75.48**	49.74	38.01	26.67	73.13	48.74	33.69
Kimi K3 High	25.05	70.23	44.89	36.43	25.05	70.40	34.72	37.41
GPT-6 Sol High	26.87	72.43	52.84	35.73	27.47	72.25	45.55	35.18
GPT-5.6 Sol High	25.86	72.31	57.69	32.48	26.46	71.93	52.71	32.24
Qwen3.5-4B	17.98	64.78	40.25	28.17	18.38	63.94	39.26	28.04
Qwen3-VL-8B Thinking	14.95	61.56	38.70	27.65	15.56	60.43	33.66	30.39
Qwen3-VL-8B Instruct	11.52	61.39	36.02	27.22	9.90	58.17	31.24	25.15

Overall performance. Table 2 compares ten VLM baselines on Base and Augmented. Doubao Seed 2.0 Pro leads Base EFC at 42.53, closely followed by Qwen3.8-Max at 42.35. On Augmented, Qwen3.8-Max leads at 41.66, followed by Qwen3.8-27B at 39.37 and Seed at 39.22. Base EFC ranges from 27.22 to 42.53. The best exact valid-set identification score is only 33.54, achieved by Qwen3.8-Max.

Valid-set identification and option discrimination. For Qwen3.8-Max, the gap between Base EFC (42.35) and EM (33.54) reflects partial credit for nonempty, correct but incomplete selections. GPT-6 Astra High has the highest Base BA (75.48), while Qwen3.8-Max has the highest EM and Seed the highest EFC. On Augmented, Qwen3.8-Max leads BA as well (74.49). Option-level discrimination, exact-set identification, and error-free coverage therefore yield different comparisons.

False selections. Base FSR ranges from 16.20 for Seed to 57.69 for GPT-5.6 Sol High. Seed also has the lowest Augmented FSR (16.25). FSR excludes omissions and must be interpreted alongside EM, BA, and EFC.

Performance on paired questions. Across ten VLM baselines, the mean absolute Base–Augmented EFC difference is 1.63 points, and eight differ by fewer than three points. Aggregate performance thus changes modestly under the paired variants. Each variant uses its own reviewed answer sets, so similar aggregate scores do not establish unchanged choices for corresponding candidates. Appendix C reports the transformation breakdown.

Option-order sensitivity. Some model rankings change under the alternate fixed option order (Appendix C.2); scores from a single ordering therefore do not establish position-invariant performance.

5.2 Option-Level Error Correction under Evidence Interventions

Response to additional support. Across three model-specific diagnostic cohorts, Perception Ref corrects 58.7–70.7% of initial ordinary-option errors. Among the cases still incorrect, adding Reasoning Ref while retaining Perception Ref corrects 61.0–75.1%. Together, the two stages correct 86.2–91.4% of initial errors, leaving 8.6–13.8% unresolved (Table 3). These rates distinguish judgments corrected with perceptual support from those corrected only under combined support; they do not identify the original error cause. The cohorts contain different initial errors and are not a head-to-head model ranking. Unresolved cases include failed outputs as well as valid but incorrect decisions.

Table 3: Correction of Base option errors under staged evidence interventions. Counts and denominator conventions are detailed in Appendix E.2.

Model	Perceptual-support recovery (%) $100P/N$	Combined-support recovery (%) $100R/(N-P)$	Judgment-stage unresolved (%) $100J/N$
Qwen3.8-Max	65.5	75.1	8.6
Qwen3.8-27B	70.7	61.0	11.4
Qwen3.5-4B	58.7	66.5	13.8

5.3 GRPO Training

Table 4: Base performance before and after GRPO training of Qwen3.5-4B.

Model	EM (%)↑	BA (%)↑	FSR (%)↓	EFC (%)↑
Qwen3.5-4B	17.98	64.78	40.25	28.17
Qwen3.5-4B + GRPO	23.43	68.34	51.92	33.37
Change	+5.45	+3.56	+11.67	+5.20

On Base, GRPO raises exact matches from 89 to 116 of 495 questions (EM: 17.98 to 23.43), BA from 64.78 to 68.34, and EFC from 28.17 to 33.37. At the same time, FSR rises from 40.25 to 51.92. Thus, exact matches become more frequent and average error-free coverage rises, while false selections increase overall. The EFC gain does not imply uniformly better rejection of unsupported descriptions. The EFC improvements are 5.20 and 4.21 points on Base under the two option orders, and 3.78 and 4.74 points on Augmented, respectively. These are separate fixed-order evaluations rather than an average across permutations (Appendix F.3). Further training details appear in Appendix F.

6 Conclusion

CoSpatial-Bench evaluates joint recognition of coexisting valid spatial descriptions within a supplied candidate set. Across ten VLM baselines on Base, the highest Base EM is 33.54 (Qwen3.8-Max), and the highest EFC is 42.53 (Doubao Seed 2.0 Pro). Paired variants show modest aggregate EFC changes across question formulations and spatial conditions, while separate Ref-assisted follow-ups characterize which initially incorrect ordinary-option decisions are corrected. Across three diagnostic cohorts, 8.6–13.8% of initial errors remain unresolved after both support stages, including failed outputs. Training Qwen3.5-4B on TCR-QAS data with GRPO improves EFC across both benchmark variants and both evaluated option orders. On Base, EFC rises from 28.17 to 33.37, while FSR rises from 40.25 to 51.92. This pattern is consistent with a less conservative selection tendency: selecting more valid descriptions on questions without false selections raises EFC, whereas accepting more incorrect options elsewhere raises FSR. The EFC gain therefore does not imply better rejection of unsupported descriptions overall. These findings support evaluating coverage and soundness together.

Ethics statement

The benchmark is intended for research on spatial relation understanding, rather than applications requiring certified geometric accuracy.

Reproducibility statement

The appendix describes dataset composition, construction and review procedures, evaluation conventions, reported training settings, parser behavior, and available prompt templates.

AI use statement

Language models assisted candidate generation, automated review, and manuscript drafting. The construction protocol identifies their roles; the authors are responsible for the analyses and claims in this paper.

References

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19129–19139, June 2022.

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

Anja Belz, Adrian Muscat, Pierre Anguill, Mouhamadou Sow, Gaétan Vincent, and Yassine Zinessabah. SpatialVOC2K: A multilingual dataset of images with annotations and features for spatial relations between objects. In Emiel Krahmer, Albert Gatt, and Martijn Goudbeek (eds.), *Proceedings of the 11th International Conference on Natural Language Generation*, pp. 140–145, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6516. URL `https://aclanthology.org/W18-6516/`.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024.

Yuchong Chen, Bowei Zou, Yuhan Chen, Yifan Fan, Xinyu Li, Shujun Cao, and Yu Hong. Looking beyond the one: Operationalizing and eliciting visual ambiguity in VLLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24306–24323, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1115. URL `https://aclanthology.org/2026.acl-long.1115/`.

Ziyang Gong, Wenhao Li, Xianzheng Ma, Songyuan Li, Zhaokai Wang, Songze Li, Jiayi Ji, Xue Yang, Gen Luo, Junchi Yan, and Rongrong Ji. SpaCE-10: A comprehensive benchmark for multimodal large language models in compositional spatial intelligence. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=Df7UjwEgIx`.

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

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=RN5KLywTll.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models, 2026. URL https://openreview.net/forum?id=mYNi6X3XWY.

Zhixing Li and Yinan Yu. From hallucination to grounding: Diagnosing visual spatial intelligence via crisp, 2026. URL https://arxiv.org/abs/2606.26535.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023.

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. Can multimodal large language models understand spatial relations? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 620–632, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.31. URL https://aclanthology.org/2025.acl-long.31/.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934, October 2025a.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 140751–140774. Curran Associates, Inc., 2025b. doi: 10.52202/085713-4703. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/cdd63c6c813a6ec826051b6a49f11b5e-Paper-Conference.pdf.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3d: Situated question answering in 3d scenes. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

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

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. https://qwen.ai/, August 2026b. Accessed: 2026-09-26.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Dynamic spatial aptitude training for multimodal language models, 2025. URL https://arxiv.org/abs/2412.07755.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21440–21455, 2024.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards comprehensive evaluation for spatial intelligence. In *Conference on Computer Vision and Pattern Recognition 2026*, 2026. URL https://openreview.net/forum?id=Yslp1qHkbI.

Yuechen Xie, Xiaoyan Zhang, Yicheng Shan, Zhu Hao, Rui Tang, Rong Wei, Mingli Song, Yuanyu Wan, and Jie Song. Spatialqa: A benchmark for evaluating spatial logical reasoning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2645–2657, June 2026.

Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *International Conference on Computer Vision (ICCV)*, 2019.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-bench: A benchmark for multi-image spatial intelligence. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gHRoX4vXm3.

Appendix Contents

Page

A Comparison with Related Benchmarks 14

B Dataset Details and Quality Control 15
B.1 Data Sources and Benchmark Composition 15
B.2 Candidate Generation and Question Assembly 16
B.3 Annotation Review 16
B.4 Training Subset and Data Isolation 16

C Augmented Benchmark Results and Examples 16
C.1 Overall Augmented Results 16
C.2 Separate Evaluation of Option Order 18
C.3 Attribute-Based Referring Expressions 19
C.4 Person-Centered Viewpoints 20
C.5 Two-Dimensional Mirror Transformations 20

D Qualitative Benchmark Examples 21
D.1 Multiple Valid Spatial Descriptions 21
D.2 Missed Valid Answers and False Selections 22
D.3 Example with No Correct Ordinary Option 22

E Staged Evidence Interventions 23
E.1 Intervention Protocol and Ref Review 23
E.2 Recovery Categories and Statistical Conventions 23

F Reinforcement Learning Details 24
F.1 Training Objective and Reward Definition 24
F.2 Reported Training Configuration 24
F.3 Post-Training Evaluation 25

G Prompts and Reproducibility 25
G.1 Evaluation Conventions 25
G.2 Prompt Templates 26
G.2.1 Candidate generation 26
G.2.2 Multi-select evaluation 27
G.2.3 Question and known-information augmentation 28
G.2.4 Perception and reasoning Ref generation 31
G.2.5 Candidate evaluation with known information 34

A Comparison with Related Benchmarks

Table 5: Task formulations (a) and diagnostic protocols (b). Entries distinguish annotations, required responses, and evidence supplied during evaluation; they do not measure isolated cognitive abilities.

(a) Task formulations

Dataset	Unit	Response / scoring	Description scope
VG (2017)	Regions, objects, pairs	Graphs, descriptions, QA pairs (annotations)	Dense scene annotations
GQA (2019)	Compositional image question	One binary or open answer	Can combine multiple objects and relations
VSR (2023)	Image–relation statement	True/false classification	One specified relation between two objects
SpatialVOC2K (2018)	Ordered object pair	Ranked prepositions; top-n hit and top-1 precision	Multiple applicable relation labels for a fixed pair
SpaCE-10 (2026)	Compositional scene question	Single choice; or 2-of-5 with both required	Options can concern different objects and multiple claims
SpatiaLQA (2026)	Procedural spatial question	Generated steps / dependencies; content and precondition precision, recall, and F1	Procedure content and ordering
CoSpatial-Bench **(Ours)**	Same-target spatial question	Multi-select: variable k; k undisclosed; EM, EFC	Supplied descriptions; potentially different anchors and multiple claims

(b) Diagnostic protocols

Study	Evidence and its use	Diagnostic observation
Spatial-MM (2024)	Bounding boxes / scene graphs supplied for spatial QA	Answer accuracy and reasoning paths
MMSI-Bench (2026)	Gold answers and reference reasoning supplied to an error-analysis judge	Error categories in model reasoning traces
CRISP (2026)	Ground-truth / predicted graphs as inputs; QA and graph generation	Cross-task consistency and effects of structural inputs
CoSpatial-Bench **(Ours)**	Initial multi-select response; perceptual Ref, then combined perceptual and reasoning Refs for remaining incorrect ordinary-option decisions	Correction or persistence under staged, cumulative Ref conditions

k denotes the number of correct ordinary options; variability and disclosure are separate properties. SpaCE-10 follows the published double-choice rule; prompt-level cardinality disclosure is not compared.

Additional distinctions. SpatiaLQA evaluates generated procedural steps and their preconditions using content and precondition precision, recall, and F1 (Xie et al., 2026), whereas *Looking Beyond the One* studies visual ambiguity across possible referents (Chen et al., 2026). These settings differ from recognizing supplied descriptions of a fixed target. SpatialScore spans spatial capabilities, modalities, and response formats (Wu et al., 2026); What'sUp changes configurations while retaining object identities, and 3DSRBench uses flips and paired common/uncommon viewpoints (Kamath et al., 2023; Ma et al., 2025a).

B Dataset Details and Quality Control

B.1 Data Sources and Benchmark Composition

CoSpatial-Bench Base contains 495 questions with 2,000 ordinary candidate answers: 1,031 correct and 969 incorrect. Appending one "None of the above options is correct" option ($\perp$) per question yields 2,495 options, of which 1,079 are correct and 1,416 are incorrect. Augmented contains 495 questions with 2,021 ordinary candidates (987 correct and 1,034 incorrect). Including $\perp$, its 2,516 options comprise 1,044 correct and 1,472 incorrect options. These counts use the reviewed v1.0 labels, including the follow-up label corrections.

The benchmark draws images from eight sources, with the distribution shown in Table 6. Augmented uses the same images and target objects, with transformed questions.

Table 6: Source composition of CoSpatial-Bench Base. Percentages use largest-remainder rounding to sum to 100.00.

Source	Questions	%	Source	Questions	%
VSR	369	74.54	Visual Genome	64	12.93
GQA	25	5.05	PerceptionBench	19	3.84
MazeBench	8	1.62	Generated Spatial Grid	6	1.21
OmniSpatial	3	0.61	MathCanvas-Instruct	1	0.20
			Total	495	100.00

Table 7 reports the distribution of the number of ordinary options m per question, excluding $\perp$. Both variants contain 495 questions, with 2,000 ordinary options in Base and 2,021 in Augmented.

Table 7: Number of ordinary options per question. Option $\perp$ is excluded.

Ordinary options m	Base questions	Augmented questions
1	14	13
2	56	52
3	118	122
4	125	123
5	105	103
6	49	49
7	23	27
8	5	6
Total questions	495	495

Let $k = |G_{\text{ord}}|$ denote the number of correct ordinary options for a question, excluding $\perp$. Table 8 groups questions by whether they contain multiple valid ordinary descriptions. This is the case for 317 of 495 Base questions (64.0%) and 302 of 495 Augmented questions (61.0%). The remaining questions have zero or one correct ordinary option.

Table 8: Grouped distribution of correct ordinary-option counts per question. Percentages use all 495 questions in each variant; option $\perp$ is excluded.

Correct ordinary options k	Base		Augmented	
	Questions	%	Questions	%
$k < 2$	178	36.0	193	39.0
$k \geq 2$	317	64.0	302	61.0
Total	495	100.0	495	100.0

B.2 Candidate Generation and Question Assembly

Human annotators select images from the eight sources and determine the corresponding questions, focusing on targets that admit multiple valid spatial descriptions. Seven vision-language models then generate candidate descriptions through rollouts for the same image and question, using different spatial anchors and levels of detail. We cluster the generated candidates by semantic similarity. Human reviewers then inspect the clusters and remove descriptions with identical or near-identical meanings. Descriptions using different reference objects or reference frames are treated as distinct relational descriptions and are not merged. Human reviewers assess candidate correctness as described below. The retained descriptions for the same image, question, and target are assembled into a multi-select instance. Each question includes a $\perp$ option, which is correct only when all ordinary candidates are incorrect.

B.3 Annotation Review

Two human reviewers jointly assessed each Base candidate as a complete claim, checking the target, reference object, coordinate frame, and spatial relation. Candidates on which the reviewers disagreed were discarded. All 2,000 ordinary candidates in the final Base set passed this review.

The training annotations follow a separate protocol. Gemini 3.1 Pro, Gemini 3.8 Flash, and GPT-5.6 Sol independently assign candidate labels; only unanimous decisions are retained. Human reviewers audit 500 questions sampled from the final 1,200-question training subset. The retained labels agree with human judgments for 85% of the candidates in these questions.

B.4 Training Subset and Data Isolation

The training pool contains 1,691 questions and a separate 95-question validation split. Questions sharing an image stay in the same split, and exact image-content checks find no overlap between training, validation, and either benchmark variant. For difficulty filtering, Qwen3.5-4B produces eight rollouts per question. Questions for which all eight rollouts are correct or all eight are incorrect are excluded, yielding the 1,200-question RL subset. These filtering conditions refer to model responses, whereas the following composition categories refer to candidate labels. The subset contains 1,025 mixed, 109 all-correct, and 66 all-incorrect questions, with 2,712 correct and 2,424 incorrect ordinary options. Each training item has at most five ordinary options plus $\perp$.

C Augmented Benchmark Results and Examples

Paired evaluation. Base and Augmented pair the same images and target objects across changes in referring expressions, viewpoints, or spatial conditions. Each version is evaluated against its own reviewed answer set, accounting for changes in candidate validity induced by the transformation.

C.1 Overall Augmented Results

Figure 5 shows ten baseline configurations on Augmented random-42, ranked by their Augmented EFC. Table 9 reports Qwen3.8-Max's performance on the 324 attribute-based, 142 person-centered, and 29 mirror-transformed questions. Group membership is aligned by question ID.

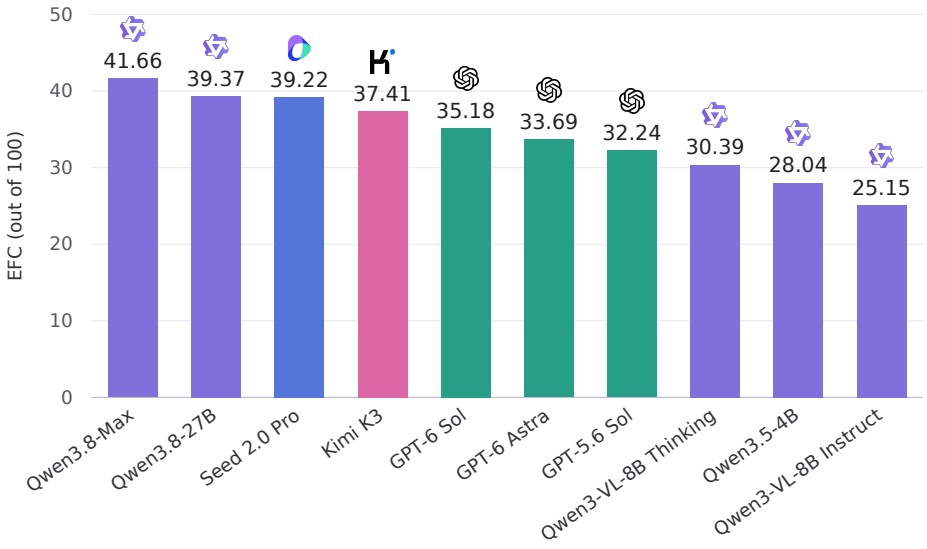

Figure 5: Augmented results on 495 questions with option-order seed 42. Qwen3.8-Max leads with EFC 41.66, followed by Qwen3.8-27B (39.37) and Doubao Seed 2.0 Pro (39.22). Models are ranked by Augmented EFC; full results are reported in Table 2.

Table 9: Transformation-specific results for Qwen3.8-Max, option-order seed 42.

Transformation	N	Base				Augmented			
		EM↑	BA↑	FSR↓	EFC↑	EM↑	BA↑	FSR↓	EFC↑
Attribute-based reference	324	30.25	73.08	**43.38**	39.40	**31.79**	**74.83**	45.73	**40.70**
Person-centered viewpoint	142	**30.99**	**73.57**	43.52	**39.91**	27.46	71.17	**28.61**	38.39
Horizontal mirror	29	**82.76**	**96.15**	**10.42**	**87.36**	58.62	87.07	18.37	68.39

Notes. N counts paired questions. All metrics follow Section 4.3. Bold indicates the better Base/Augmented value within a row, including ties. FSR denominators (Base, Augmented) are $(597, 597)$, $(324, 388)$, and $(48, 49)$ in row order.

Stability across benchmark variants. For random-42, eight of the ten baselines differ by less than three EFC points across variants. Seed leads Base, while Qwen3.8-Max leads Augmented. The top three models are Seed, Qwen3.8-Max, and Qwen3.8-27B in both variants, but their ordering changes. Aggregate similarity does not imply that individual predictions are invariant under transformations.

Sensitivity by transformation. Qwen3.8-Max's EFC changes from 39.40 to 40.70 for attribute-based references and from 39.91 to 38.39 for person-centered viewpoints. The mirror group shows a larger decline, from 87.36 to 68.39, but contains only 29 questions. Different groups have different starting difficulty and valid answer sets; these changes do not isolate a causal effect of reference-frame reasoning.

C.2 Separate Evaluation of Option Order

Table 10: results with option-order seed 24.

Model	Base (%)				Augmented (%)			
	EM↑	BA↑	FSR↓	EFC↑	EM↑	BA↑	FSR↓	EFC↑
Doubao Seed 2.0 Pro High	26.46	64.09	**16.82**	40.82	26.06	65.98	**16.15**	**40.94**
Qwen3.8-Max	**33.33**	74.66	43.65	**41.46**	**29.90**	**74.19**	38.10	40.42
Qwen3.8-27B	28.08	71.36	39.32	38.78	24.04	69.89	31.24	39.02
GPT-6 Astra High	30.91	**75.49**	50.36	38.07	27.27	73.45	48.74	34.61
Kimi K3 High	25.66	71.90	43.03	36.46	24.85	69.70	37.91	35.81
GPT-6 Sol High	25.86	71.85	53.66	33.62	26.06	71.65	47.39	33.86
GPT-5.6 Sol High	25.25	71.65	58.51	31.16	26.26	72.54	52.71	32.96
Qwen3.5-4B	20.61	66.34	40.35	28.90	18.18	63.14	38.97	26.15
Qwen3-VL-8B Thinking	16.16	61.64	41.38	29.14	13.94	60.41	34.82	28.56
Qwen3-VL-8B Instruct	10.71	60.97	38.39	25.47	10.10	58.43	31.91	26.44

Notes. Scores use a 0–100 scale. Observed runs use all 495 questions, including failed outputs. FSR excludes ⊥, with denominators 969 (Base) and 1,034 (Augmented). Bold marks the best available value in each column. Rows follow Base random-42 EFC. Option orders are reported separately without averaging.

The random-24 permutation changes option order while preserving the questions and candidate labels within each variant. It is reported separately from random-42, without averaging or choosing a model's better order. On random-24, Qwen3.8-Max leads Base at 41.46, followed by Seed at 40.82. On Augmented, Seed leads at 40.94, followed by Qwen3.8-Max at 40.42 and Qwen3.8-27B at 39.02. Some models exchange positions across orders, so a single order does not establish order-invariant performance. These two evaluations are different fixed option arrangements, not repeated samples of one arrangement.

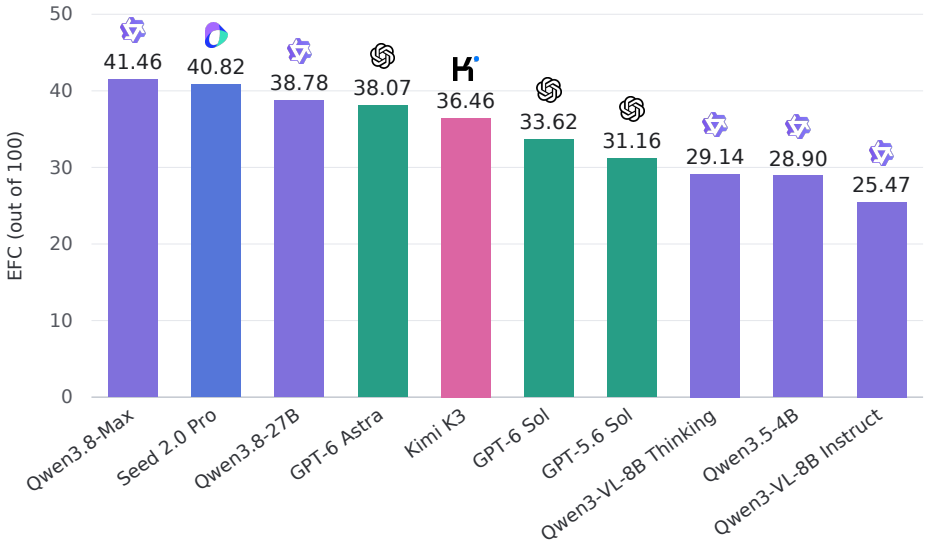

Figure 6: Option-order ablation on Base with seed 24. Models are ranked by Base seed-24 EFC; scores are not averaged across orders.

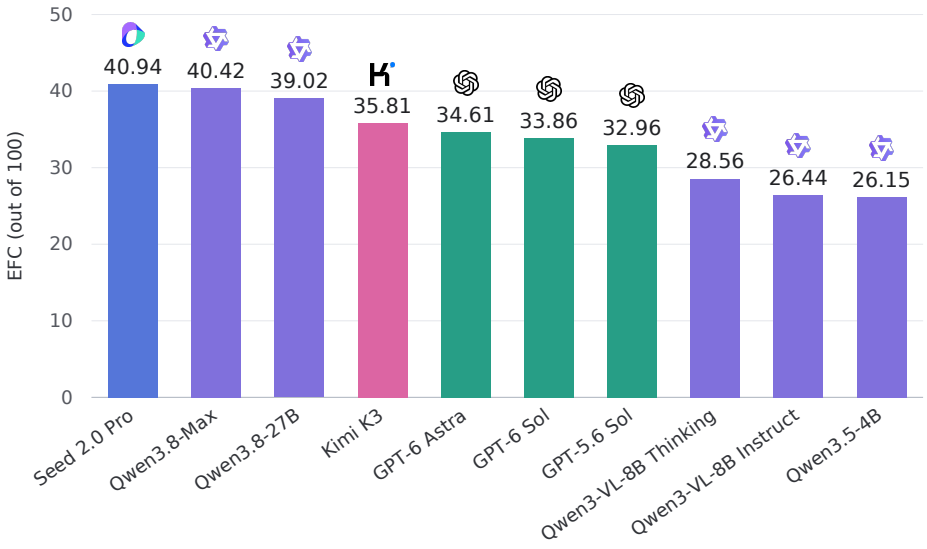

Figure 7: Option-order ablation on Augmented with seed 24. Models are ranked by Augmented seed-24 EFC. All reported metrics appear in Table 10.

C.3 Attribute-Based Referring Expressions

Attribute-based rewriting replaces an object name with a description of its function or properties. The description must uniquely identify the original target and preserve the validity of its candidate answers. Figure 8 illustrates this transformation for a television: the question refers to its function, while the options and gold answer set remain unchanged.



Figure 8: Attribute-based referring-expression example. The rewrite changes how the television is identified without changing the spatial task.

Base: Where is the television?
Augmented: Where is the device used to display or play audiovisual content?
Options:
A. To the right of the fireplace;
B. Beside the small green table;
C. None of the above options is correct.
Gold answer set: {A, B} in both versions.

C.4 Person-Centered Viewpoints

Person-centered questions use a depicted person's body orientation as the reference frame. Changing the frame can change which spatial relations hold. Reviewers therefore check that the person's orientation is discernible and reassess descriptions that depend on it. Figure 9 shows how the question identifies the reference person and defines the forward direction.



Figure 9: Person-centered viewpoint example. The augmented question specifies the reference person's body-centered frame. Both versions use the same options and have gold answer set $\{B, C\}$.

Base: Where is the child wearing green?

Augmented: Taking the current body-facing direction of the boy in the gray short-sleeved T-shirt seated at the far end of the dining table as "forward," where is the child wearing green from that boy's perspective?

Options (both versions):

A. In a seat on the window side of the dining table, directly facing the tabletop with a green plate and a beer bottle.

B. The child wearing green sits beside the dining table, near the woman wearing a burgundy top.

C. Seated on the right of the side of the dining table closest to the plants.

D. At the right end of the long dining table.

E. None of the above options is correct.

Gold answer set: $\{B, C\}$ in both versions.

C.5 Two-Dimensional Mirror Transformations

For eligible two-dimensional scenes, the question asks the model to consider a horizontal reflection of the spatial layout. The displayed image remains unchanged, and the coordinate system and its labels stay fixed. Directional descriptions are transformed under the same convention and reviewed. In Figure 10, "right" becomes "left" in option A, preserving its validity under the reflected layout.

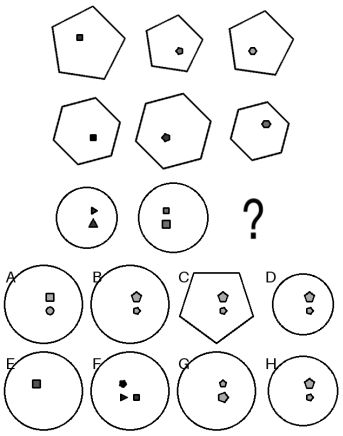

Figure 10: Horizontal-mirror example. The image remains unchanged; the augmented question specifies a counterfactual reflection. Option A changes from right to left, while options B–D and the gold answer set are unchanged.

Base: Where is the question-mark shape?

Augmented: Suppose all spatial elements in the image (the shapes and question mark in the upper matrix, and the option shapes below) are reflected left to right about the image's vertical centerline, while the matrix row and column indices, the option letters, and all other coordinate definitions remain fixed. Where is the question-mark shape?

Base option A: To the right of the circle containing two small squares in the third row.

Augmented option A: To the left of the circle containing two small squares in the third row.

Options B–D (both versions):

B. Slightly below the center of the image.

C. At the very top of the options area.

D. None of the above options is correct.

Gold answer set: $\{A\}$ in both versions.

D Qualitative Benchmark Examples

The following examples illustrate complete and partial valid-set identification, incorrect selections, and $\perp$ answers.

D.1 Multiple Valid Spatial Descriptions



Figure 11: A coexisting-relation example. A and B use different anchors to describe the same television; C introduces an unsupported relation and D is false because valid ordinary options exist.

Question: Where is the television?

Options:

A. Directly above the shelf filled with DVDs.
B. On the row of bookshelves against the wall to the right of the bicycle, above the rows of books.
C. On the top shelf, directly below the picture frame.
D. None of the above options is correct.

Gold answer set: {A, B}.

D.2 Missed Valid Answers and False Selections

A missed valid answer is a gold option omitted from the predicted set. A false selection is a selected option outside the gold answer set. These errors distinguish incomplete recognition of valid descriptions from acceptance of invalid descriptions.

Table 11 compares current Base random-42 predictions for the question in Figure 11. GPT-6 Astra High selects both correct options. Qwen3.8-Max selects only B and receives half credit. Qwen3.8-27B selects B and incorrect option C, receiving zero credit.

Table 11: Three random-42 predictions for the same television question.

Model	Prediction	EFC contribution	Outcome
GPT-6 Astra High	$\{A, B\}$	1.0	Exact identification
Qwen3.8-Max	$\{B\}$	0.5	Omitted A
Qwen3.8-27B	$\{B, C\}$	0.0	False C; omitted A

D.3 Example with No Correct Ordinary Option

Figure 12 shows a Base question in which no ordinary option is correct, making "None of the above options is correct" the sole correct answer.



Figure 12: A Base example in which "None of the above options is correct" is the only correct option.

Question: Where is the white spotted dog?

Options:

A. To the right of the red boat, near the shore;
B. None of the above options is correct.

Gold answer set: $\{B\}$.

E Staged Evidence Interventions

E.1 Intervention Protocol and Ref Review

The initial condition C_0 is the model's first whole-question multi-select response. For each ordinary candidate, selection of an incorrect option is a false selection, while omission of a correct option is a missed valid description; both enter the initial error cohort. The corresponding prompt appears in Appendix G.2.2.

For these initially incorrect option decisions, the model first receives a Perception Ref in a single-candidate prompt (C_P). Cases still incorrect at C_P then receive the same Perception Ref together with a Reasoning Ref (C_{PR}). The image, question, candidate, and ordinary-option selection criterion are retained, and each condition uses a fresh context. The second stage therefore accumulates the two types of information rather than replacing the perceptual information.

GPT-6 Astra with high reasoning effort receives the human correctness label when generating the perceptual and reasoning information. The generation prompt appears in Appendix G.2.4. Human reviewers check the generated information only when the generator's conclusion disagrees with that label. Both types of Ref are required for inclusion in the diagnostic cohort. The Refs supplied to the evaluated model state neither the candidate's correctness label nor the gold answer set.

E.2 Recovery Categories and Statistical Conventions

Let $\mathcal{E}$ denote the set of ordinary-candidate decisions that disagree with the reviewed labels when selection or nonselection is read from the initial `selected_options` list, and for which both types of intervention information are available. Let c_i^P indicate whether the single-candidate decision under Perception Ref matches the reviewed label. For cases with $c_i^P = 0$, let c_i^{PR} indicate the same under the combined Perception and Reasoning Refs. A condition that does not produce a valid correct binary decision has indicator value zero.

We partition this diagnostic set by the first successful correction stage:

$$\begin{aligned}\mathcal{E}_P &= \{i \in \mathcal{E} : c_i^P = 1\},\\ \mathcal{E}_R &= \{i \in \mathcal{E} : c_i^P = 0,\ c_i^{PR} = 1\},\\ \mathcal{E}_J &= \{i \in \mathcal{E} : c_i^P = 0,\ c_i^{PR} = 0\}.\end{aligned} \tag{8}$$

Only candidates not corrected under the Perception Ref condition receive the combined Perception and Reasoning Ref intervention. We report the stage-specific rates

$$\begin{aligned}p_P &= 100\frac{|\mathcal{E}_P|}{|\mathcal{E}|},\\ p_R &= 100\frac{|\mathcal{E}_R|}{|\mathcal{E}| - |\mathcal{E}_P|},\\ p_J &= 100\frac{|\mathcal{E}_J|}{|\mathcal{E}|}.\end{aligned} \tag{9}$$

Thus p_R measures recovery among cases remaining after the perceptual stage, whereas p_J is the share of initial errors unresolved after both stages. The three rates have different denominators and are not additive. This candidate-level diagnostic cohort is distinct from the 495-question denominator used for EM, BA, and EFC. A rate is undefined when its denominator is zero.

For Table 3, let $N = |\mathcal{E}|$, $P = |\mathcal{E}_P|$, $R = |\mathcal{E}_R|$, and $J = |\mathcal{E}_J| = N - P - R$. In table order, the counts $(N; P, R, J)$ are $(605; 396, 157, 52)$, $(666; 471, 119, 76)$, and $(831; 488, 228, 115)$. The judgment-stage unresolved category is the residual after both Ref conditions and includes failed outputs as well as valid incorrect decisions. For Qwen3.5-4B, 57 of the 115 unresolved cases are second-stage parse failures; the other 58 yield valid but incorrect decisions. Its C_{PR} condition uses the same Perception Ref together with a Reasoning Ref; earlier Reasoning-Ref-only results are excluded. All three diagnostic cohorts come from initial errors in the Base random-42 evaluations reported in Table 2.

F Reinforcement Learning Details

F.1 Training Objective and Reward Definition

Let $\mathcal{O}_i$ denote the option set for question i, including $\perp$. The gold answer set G_i is nonempty: it contains all correct ordinary options when any exist, and otherwise contains only $\perp$.

For a valid predicted set $S_i \subseteq \mathcal{O}_i$, the per-question score is

$$s_i = \begin{cases}\dfrac{|S_i|}{|G_i|}, & \varnothing \neq S_i \subseteq G_i,\\ 0, & \text{otherwise.}\end{cases} \tag{10}$$

Invalid outputs and exhausted request failures receive zero. Selecting any incorrect option therefore yields zero, whereas selecting a nonempty strict subset of the gold answer set receives partial credit.

The RL outcome reward is

$$r_i = s_i^2. \tag{11}$$

Squaring increases the reward difference between complete valid-set identification and correct but incomplete selections. GRPO uses this reward.

F.2 Reported Training Configuration

Our GRPO configuration trains Qwen3.5-4B with eight rollouts per question, a learning rate of 10^{-6}, batch size 32, and KL-loss coefficient 0.01. Sampling uses temperature 1.0 and top-p 0.95. The maximum prompt and response lengths are 3,120 and 16,384 tokens, respectively. The run uses eight A100 40GB GPUs, seed 42, and a budget of at most 300

Table 12: Score and reward examples for $G = \{A, B\}$, where C is incorrect. Here s is the per-question EFC contribution before scaling to 0–100.

Prediction	Score s	Exact match	RL reward s^2
$\{A\}$	0.5	0	0.25
$\{A, B\}$	1	1	1
$\{A, C\}$	0	0	0
$\varnothing$	0	0	0
Invalid output	0	0	0

steps or six epochs, with validation and checkpointing every 25 steps. An independent 95-question training-domain split is used for validation. The evaluation baseline uses all 495 Base questions; Table 4 reports the main Base comparison; Table 13 reports EFC across both variants and option orders.

F.3 Post-Training Evaluation

We evaluate the baseline and GRPO-trained checkpoint on Base and Augmented using the same questions, reviewed answer sets, option orders, and scoring rules. Table 4 reports the main Base comparison for option-order seed 42. Table 13 reports EFC for seeds 42 and 24 separately, with improvements in all four settings. The examples in Table 12 illustrate the scoring and reward functions.

Table 13: EFC (%) before and after GRPO training of Qwen3.5-4B across benchmark variants and option orders.

Model	Base 42 (%)	Base 24 (%)	Augmented 42 (%)	Augmented 24 (%)
Qwen3.5-4B	28.17	28.90	28.04	26.15
Qwen3.5-4B + GRPO	33.37	33.11	31.82	30.89
Difference	+5.20	+4.21	+3.78	+4.74

G Prompts and Reproducibility

G.1 Evaluation Conventions

The EM, BA, and EFC definitions in Section 4.3 average over all 495 questions and include $\perp$ in the option set. A valid empty selection receives zero EM and EFC and a BA of 50; an invalid response receives zero for all three metrics. Selecting $\perp$ together with an ordinary option is invalid.

FSR counts incorrect ordinary options accepted as correct, using 969 Base and 1,034 Augmented options as its respective denominators. It excludes $\perp$. Recorded candidate decisions are counted individually; a failure elsewhere in the same question does not remove them. Empty or failed responses contribute no false selections but remain in the denominator.

Models return their selections in a JSON object with a `selected_options` list. Each option must appear at most once; unknown options, duplicate entries, malformed JSON, and extra fields are treated as invalid responses. An empty list denotes no selection, not $\perp$.

Models and option order. We compare ten baseline VLM settings: Qwen3.8-Max, Qwen3.8-27B, Qwen3.5-4B, Qwen3-VL-8B Thinking, Qwen3-VL-8B Instruct (Qwen Team, 2026b;a; Bai et al., 2025), Doubao Seed 2.0 Pro, Kimi K3, GPT-6 Astra, GPT-6 Sol, and GPT-5.6 Sol. The latter five use High configurations. We also evaluate a trained Qwen3.5-4B + GRPO checkpoint. The v1.0 release uses two fixed permutations of ordinary options, identified by seeds 42 and 24. Main results and diagnostic cohorts use random-42; random-24 is reported separately in Appendix C.2. Scores are not averaged across orders. Each variant

and order has the same 495 question IDs, with its own supplied options and reviewed answer sets.

For Doubao Seed 2.0 Pro on Augmented random-24, 484 of 495 requests succeeded; the 11 failures remain in the denominator. Request and parse failures are retained in the 495-question denominator for every setting. Recovery and retry policies differ across models.

Diagnostic evaluation uses Base random-42 errors and fresh single-candidate requests. The 4B combined-Ref condition uses thinking mode, temperature 1, and a 32,768-token response limit.

Qualitative construction examples use their displayed option letters; quantitative comparisons use the released random-42 or random-24 ordering.

G.2 Prompt Templates

The intervention prompt follows the same evaluation rules for both intervention stages.

G.2.1 Candidate Answer Generation

System Prompt

```
You are a visual question answering model. Examine the image and answer the
user's question naturally and accurately.

You may provide several natural answers that describe the same location using
different references, up to a maximum of 4. For real-world scenes, prioritize
descriptions based on the actual spaces, objects, or structures in the scene,
such as which shelf, tabletop, or room the target is in or on, or which object
it is next to; do not prioritize image coordinates such as ‘on the left side of
the image’ or ‘in the middle of the image.’ For maps, chessboards, flowcharts,
geometric diagrams, and similar images, answer naturally using the image's own
structure and labels.

All answers should refer to the same target instance and describe the same
location, differing only in their references; do not switch to other objects
of the same type across answers. Provide as many answers as can be expressed
naturally. Do not force additional answers to reach a particular number, and
do not merely rephrase them with synonyms.

Return only a valid JSON object:
{
  "answers": [
    {
      "answer": "A concise answer",
      "reasoning": "A brief explanation of the evidence in the image used to
      locate the target"
    }
  ]
}

answers must contain at least 1 and at most 4 items, each containing only
answer and reasoning. Do not use Markdown code fences or output any text
outside the JSON. Both answer and reasoning must be strings.
```

User Prompt

```
{question}
```

G.2.2 Multi-Select Evaluation

System Prompt

```
Based on the image and the question, select all correct answers. There may be
one or more correct options.

Judge each ordinary option independently and select every correct option.
Different options may use different reference objects, wording, or valid
paths; you are not required to choose a single best answer. Option text is a
description to be verified, not evidence establishing that another option is
true, and must not supply content missing from another option.

Verify the target, attributes, reference objects, and spatial relations in
each option, examining supporting and conflicting evidence using the image.
Select an option when all its key descriptions are supported and it can answer
the original question. For location questions, a description should
effectively narrow down the target's location; it need not pinpoint a unique
location, and neither answer length nor whether every detail is included is a
criterion. Do not exclude an option merely because its wording is unfamiliar,
its reference objects differ, or it requires further spatial reasoning. Do not
select an option if it contains a key contradiction, does not meaningfully
locate the target, or still lacks evidence supporting a key description after
verification.

Spatial-relation conventions:
1. First follow any perspective and coordinate requirements explicitly
specified by the question, then determine the coordinate system used by the
current option.
2. Subject to the question's constraints, use image coordinates when the
option explicitly specifies image or picture directions.
3. When the option does not explicitly specify a coordinate system, use the
reference object's own coordinate system if its intrinsic orientation can be
established; otherwise, prioritize clear scene structure. Use natural
image-based relations only when neither the reference object's own coordinate
system nor clear scene structure applies.
4. If the reference object's orientation or coordinate system cannot be
reliably determined, do not force confirmation of directional relations that
depend on it, and do not invent facts that cannot be confirmed in the image.
5. Judge front/back and depth relations using orientation, occlusion,
perspective, contact, and scene structure together, not solely from
two-dimensional vertical positions.
6. Relations with strong qualifiers such as "directly," "entirely," or
"tightly against" require more direct image evidence than ordinary relations.

Select the final option, "None of the above options is correct", only when you
have verified each ordinary option and judged all of them incorrect. It must
not be selected together with any other option. Failure to confirm that an
option is correct does not establish that it is incorrect; do not select the
final option merely because you are unsure. If all ordinary options are
correct, select all ordinary options. After verification, if you can neither
select a correct option nor judge all ordinary options incorrect, return an
empty list.

Output only valid JSON: {"selected_options": ["option ID"]}. List every
selected ID in option order, without duplicates, explanations, or any other
text.
```

User Prompt

```
Question:
{question}

Options:
{options}
```

G.2.3 Question and Known-Information Augmentation

System Prompt

```
You are a "question and reference answer rewriter" for a spatial understanding
benchmark, not a candidate answer generator or a label classifier.

Goal: Given an image, an original question, and an original reference answer,
generate exactly one augmented question. Select only one augmentation type for
each image. Candidate answers will not be provided to you. For mirrored
samples, the program will apply the corresponding deterministic transformation
to the candidate answers after you return your output. You must not infer,
supply, or output candidate answers.

I. Augmentation types and selection priority

1. horizontal_mirror_2d (highest priority)
Use this for structured 2D scenes, including but not limited to coordinate
grids, chessboards, mazes, maps, flowcharts, geometric diagrams, and manually
arranged 2D schematics. Ordinary photographs do not belong to this category
merely because they consist of 2D pixels.

Rewriting rule: Prepend an explicit counterfactual condition to the original
question: reflect the spatial layout elements in the scene (including objects,
game pieces, the start, the destination, and obstacles) horizontally about the
image's vertical centerline, while keeping the background grid, coordinate
origin, axis directions, row and column indices, board labels, and all other
coordinate definitions unchanged. Do not modify the image itself.

The reference answer must be changed to the correct answer under the mirrored
condition. Observe the following rules:
- Swap left and right, east and west, northeast and northwest, and southeast
and southwest;
- A horizontal reflection does not affect up/down, north/south, front/back,
inside/outside, contact, overlap, near/far, or similar relations;
- If there are W horizontal positions indexed from zero, (x, y) becomes
(W-1-x, y);
- Reflect board column labels according to the board width. For an
eight-column board, for example, swap a and h, b and g, c and f, and d and e;
- For maze paths, swap L and R, while leaving U and D unchanged; "unreachable"
remains "unreachable";
- Both the target and the reference object are reflected, so their relative
relationship must be reassessed.

2. person_body_viewpoint
Use this only for unstructured 2D scenes containing a person who can be
uniquely described and whose body orientation is sufficiently clear. Do not
select the person who is the target of the question as the viewpoint person.
Do not use this type if the person cannot be uniquely identified, their body
orientation is unclear, or the reference answer cannot be reliably converted.

Use the following question format:
"Taking the current body orientation of {unique visual description of the
person} in the image as 'forward', from this person's viewpoint, {original
question}"

Here, "viewpoint" refers to the person's body coordinate frame, not an
assessment of visibility. The direction the person's body faces defines front,
behind their body is back, their left-hand side is left, and their right-hand
side is right. Even a target behind the person may be described as "behind the
person". The reference answer must also be converted to this person's body
coordinate frame, while retaining any nondirectional location information that
remains valid. Do not directly equate image left/right with the person's
left/right. Reference answers of this type always require human confirmation.

3. subject_attribute
```

System Prompt (continued)

```
Use this for the remaining images. Replace only the name of the target subject
in the original question with an indirect description that requires one step
of semantic, functional, or commonsense reasoning to recover the original
object. The goal is for the respondent to first infer which type of object the
description refers to and then answer where it is. Simply replacing the
original object's name with a more detailed description of its appearance is
not allowed.

Prefer the following attributes:
- Function or use, such as replacing "television" with "a device for
displaying or playing audiovisual content";
- Nutritional, compositional, or typical knowledge-based attributes, such as
replacing "carrot" with "the vegetable with the highest vitamin A content"
when this description uniquely identifies it among the candidate vegetables in
the image;
- Typical capabilities, uses, superordinate categories, or stable commonsense
attributes;
- When necessary, combine two nonpositional attributes, but keep the
description natural and concise.

The following rewrites are prohibited:
- Do not rely primarily on directly visible appearance, such as color, shape,
size, age, material, packaging text, or clothing. For example, do not replace
"television" with "that old CRT television with a silver-gray casing";
- Do not replace the target name with a description that still directly
contains the original name, such as "the black television" or "the only
television in the image";
- Do not use a direct synonym, translation, brand, or model of the original
object name that makes the answer identifiable without reasoning;
- Do not construct the description using the target's location, its spatial
relationships with surrounding objects, or the content of the reference
answer;
- Do not introduce attributes that cannot be verified from the image and are
unstable or disputed as commonsense knowledge.

This type must satisfy all of the following:
- Modify only the question;
- rewritten_reference_answer must be identical to the original reference
answer, character for character;
- Keep the target instance unchanged;
- Among all visible objects in the current image, the indirect description
must refer only to the original target;
- target_rewritten must not directly contain target_original;
- decision_reason must explain "why the indirect attribute corresponds to the
original target" and "why it is unique in the current image";
- If a unique, natural description cannot be constructed, return unresolved
rather than forcing one.

Positive examples:
- "Where is the carrot?" → "Where is the vegetable with the highest vitamin A
content in the image?"
- "Where is the television?" → "Where is the device for displaying or playing
audiovisual content in the image?"

Negative examples:
- "Where is the television?" → "Where is that old CRT television with a
silver-gray casing?" (merely adds appearance details)
- "Where is the popcorn?" → "Where is the bag of popcorn with the word Spicy
printed on its packaging?" (retains the original name and relies on packaging
text)

4. unresolved
Use this only when none of the three types above can be safely applied. You
must still explain the reason and keep the question and reference answer
unchanged. The sample will be handled by a human afterward.
```

System Prompt (continued)

```
II. General constraints

- Select exactly one augmentation_type for each image.
- The question must actually be rewritten. Except for unresolved,
rewritten_question must differ from the original question.
- Do not answer the original question, generate candidate answers, or assign
labels to candidate answers.
- Do not change the target instance or rely on people, objects, or coordinate
definitions that are absent from the image.
- Return the original question and original reference answer fields unchanged.
- Explicitly describe every change to the reference in
reference_change_summary.
- Output one valid JSON object. Do not use Markdown code fences or output any
text outside the JSON.

III. Strict output structure

{
  "sample_id": "Exactly as provided in the input",
  "augmentation_type": "horizontal_mirror_2d | person_body_viewpoint |
  subject_attribute | unresolved",
  "decision_reason": "Why this augmentation type is appropriate for the image;
  for subject_attribute, also explain the reasoning from the indirect
  attribute to the original target and its uniqueness within the image",
  "original_question": "Return unchanged, character for character",
  "rewritten_question": "The single augmented question",
  "original_reference_answer": "Return unchanged, character for character",
  "rewritten_reference_answer": "The augmented Ref; for subject_attribute and
  unresolved, it must remain unchanged, character for character",
  "reference_changed": true,
  "reference_change_method": "deterministic_rule | visual_reasoning |
  unchanged | unresolved",
  "reference_change_summary": "Specify which directions, coordinates, or
  reference relationships were changed; if unchanged, state that no changes
  were made",
  "question_requires_human_review": true,
  "reference_requires_human_review": true,
  "target_original": "The original target subject; null if it cannot be
  reliably extracted",
  "target_rewritten": "For subject_attribute, the indirect semantic or
  commonsense description that does not rely on appearance and must not
  contain the original target name; null for other types",
  "viewpoint_person": "The description of the person used for
  person_body_viewpoint; null for other types",
  "orientation_evidence": "Evidence for the person's body orientation; null
  for other types",
  "confidence": 0.0,
  "warnings": []
}

Boolean values must be JSON booleans, and confidence must be between 0 and 1.
All fields must be present.
```

User Prompt

```
Examine the image provided with the message and rewrite the following
image-level question and reference answer according to the system rules.

Processing constraints supplied by the program:
1. If routing_constraints.forced_augmentation_type in the input JSON is not
null, you must use that augmentation type and must not select another one.
```

User Prompt (continued)

```
2. If routing_constraints.deterministic_reference_hint is not null, this field
contains the mirrored reference answer computed by the program using
deterministic rules. rewritten_reference_answer must match this field exactly,
character for character.
3. If the fields above are null, the program cannot determine the result, and
you must make the decision based on the image and the system rules.
4. Candidate answers will not be provided to you. For mirrored samples, the
program will transform the candidate answers using the same reflection rules
after you return your output. Do not generate, infer, or output candidate
answers.

Input JSON:
{input_json}

Return only one JSON object that strictly follows the structure specified by
the system.
```

G.2.4 Perception and Reasoning Ref Generation

The following prompt generates both Refs in one call. The input label is the human-reviewed Base candidate label. The response is constrained by a JSON schema with four top-level fields: `perception_reference`, `reasoning_reference`, `given_label`, and `label_consistency`. The generated Refs supplied to evaluated models exclude the label, consistency check, and generation-side audit fields; Appendix E.1 describes when human review is triggered.

System Prompt

```
[Role]

You are a high-reliability generator of visual evidence and spatial-reasoning
references for a spatial-location understanding experiment.

The input includes an image, a question, one ordinary candidate answer, and
the candidate's confirmed ground-truth label. This prompt generates
intermediate information for a later attribution experiment; it does not run
that intervention. The later experiment may treat a Perception Ref or a
Reasoning Ref as a trusted premise. Neither Ref is the original benchmark
reference answer, the model's earlier selection, or a final label.
Generate in this strict order: first record perceptual facts, then derive
spatial relations, and only then compare the result with the given label. The
label is used only for the final consistency check; it is not image evidence
and must not determine which facts or relations are written.

Generate independently from the image:

1. Perception Ref: directly observable visual facts relevant to this question
and candidate answer.
2. Reasoning Ref: how those facts imply the spatial relations relevant to the
candidate answer.

Do not invent image evidence to fit the label. If independent evidence
disagrees with the given label, set label_consistency.status to "conflict" and
explain the discrepancy without changing given_label. Descriptive fields of
either Ref must not contain the ground-truth label, an earlier model
selection, or a final judgment on the candidate. relation_status, given_label,
and label_consistency are generation-side audit fields; never pass them to the
evaluated model as a known Ref.

[Input]

The user message supplies the question, candidate answer, and confirmed label.
The corresponding image is supplied separately.
```

System Prompt (continued)

```
The label is one of:

- correct: The key descriptions are supported and answer the question. For a
location question, they must meaningfully narrow down the target's location,
but need not identify a unique position or include every detail.
- incorrect: There is a key contradiction, the candidate fails to answer the
question or meaningfully locate the target, or a key description remains
unsupported by the available image evidence. Insufficient evidence is not the
same as a visual contradiction.

[Stage 1: Perception Ref]

1. Use the question to identify the target instance.
2. Use the candidate answer to identify the reference objects, people, scene
structures, and visual attributes that need to be checked.
3. Record only the minimum relevant visual facts; omit unrelated details.
4. Every field of target, anchors, facts, image_location, and limitations must
contain only directly visible two-dimensional facts: image regions,
appearance, counts, visible outline overlap or occlusion cues, and visible
body or object orientation. image_location denotes a region in the image, not
actual front/back or depth.
5. Do not assume that a fact holds merely because the candidate answer or
given label mentions it.
6. Check the target, anchors, colors, categories, counts, visibility, visible
contact or occlusion cues, and visible body orientation. Do not turn these
cues directly into three-dimensional relations.
7. Mark a target, anchor, or attribute that cannot be identified reliably as
uncertain.
8. Do not invent exact pixel positions, occluded areas, three-dimensional
distances, orientations that cannot be discerned in the image, or motion
states. Visible body orientation is perception; deriving a person's
left/right/front/back from that orientation is reasoning.
9. Assign each perceptual fact a unique ID: F1, F2, F3, and so on.
10. No field in the Perception Ref may state whether the candidate is correct
or incorrect, whether a candidate claim is supported or contradicted, a rule
mapping 2D cues to 3D relations, a perspective transformation, or another
spatial inference. A final spatial relation supplied by a human also belongs
in reasoning, not perception.
11. Do not add meta-commentary such as "the human-confirmed relation is not an
independent image measurement." Record a brief limitation only when it
concerns an actual visible feature; otherwise use an empty string for
limitation or limitations.

[Stage 2: Reasoning Ref]

1. Split the candidate answer into key claims that affect selection. Analyze
only this candidate, without borrowing from other options or treating wording
differences from a Ref as errors.
2. Follow any perspective or coordinate requirements explicitly stated in the
question first. Then use image coordinates if the candidate explicitly
specifies them. Otherwise, prefer the reference object's own frame when its
orientation is established, followed by clear scene structure, and use natural
image coordinates only as a fallback. If a necessary frame cannot be
determined, use unresolved. Select a frame separately for each key claim:
   - image: viewer-facing image coordinates;
   - object: the reference object's intrinsic orientation;
   - scene: a road, room, shelf, board, or other scene structure;
   - not_applicable: no directional coordinate frame is involved;
   - unresolved: the required frame cannot be established reliably.
3. Cite the relevant perceptual fact IDs in every inference step.
```

System Prompt (continued)

```
4. State which perceptual facts imply which actual spatial relations. Depth,
three-dimensional relations, transformations from visible body orientation to a
person's viewpoint, and interpretation of 2D cues belong here.
reference_description should give a grounded description of the actual
location, not merely repeat the candidate answer.
5. Check supporting evidence, conflicting evidence, and insufficient evidence.
Mark each relation's audit status as:
   - supported: the visual evidence supports the relation;
   - contradicted: the visual evidence clearly refutes it;
   - not_verifiable: the image is insufficient to confirm or refute it.
6. Descriptions using different reference objects can both be valid. Do not
infer a conflict merely from different anchors or wording.
7. Do not tighten "below" into "directly below" or "beside" into "touching."
8. Infer front/back and depth using visible occlusion, perspective,
orientation, contact, and scene structure together; do not rely only on
vertical position in the image. Do not force a coordinate transformation when
the known frame does not apply.
9. A short or broad candidate answer is not wrong merely because it omits some
position details.
10. not_verifiable means insufficient evidence; do not turn it into
contradicted. Nor should an answer be considered correct merely because it is
not refuted: it must answer the question, meaningfully locate the target when
appropriate, and have support for its key descriptions.
11. relation_status, relation_evidence, and candidate_meaning are
generation-side analysis fields. A Reasoning Ref provided in a later
intervention must contain only trusted spatial relations, their coordinate
basis, and the derivation from F facts. It must not include
supported/contradicted, the ground-truth label, or a decision that the
candidate should or should not be selected. Leave limitations empty when there
is no genuine evidential limitation; do not add generic "uncertain" or "cannot
measure" boilerplate.

[Label consistency check]

Compare the independently generated evidence with the given label:

- label=correct, all key descriptions are supported, and the candidate answers
the question (meaningfully narrowing the target's location when applicable):
label_consistency.status="consistent".
- label=incorrect, and there is a key contradiction, a failure to answer or
meaningfully locate the target, or a key claim remains not_verifiable:
label_consistency.status="consistent".
- Otherwise: label_consistency.status="conflict".

This check flags possible label issues; it never changes the input label. When
incorrect results from missing support or inadequate localization, explain
that in label_consistency.reason rather than inventing a contradicted claim.
If a correct label has a key contradiction or lacks support, mark a conflict.

[Output]

Return only valid JSON conforming to the supplied JSON Schema, with no
Markdown, code fence, or additional text. The top-level keys are
perception_reference, reasoning_reference, given_label, and label_consistency.
Generation-side audit fields must not be included in a known Ref passed to the
evaluated model. The caller must separately supply
PERCEPTION_REASONING_GENERATION_SCHEMA to constrain the response.
```

User Prompt

```
[Input]

Question:
{question}

Candidate answer:
{candidate_answer}

Confirmed label:
{label}

The corresponding image is supplied separately.
```

G.2.5 Single-Candidate Diagnosis with Staged Information Support

The initial C_0 decision is read from the multi-select `selected_options` response using the prompt in Appendix G.2.2; there is no separate single-candidate C_0 prompt. The following two templates evaluate only initially incorrect ordinary options, first with a Perception Ref and then, if needed, with both the same Perception Ref and a Reasoning Ref.

G.2.5.1 Perception Ref (perception_ref).

System Prompt

```
[Prefix]
You are a neutral visual judge for a spatial-location understanding benchmark.
The candidate answer is a statement to be verified; do not presume it correct
or incorrect before verification is complete. Your task is to examine
supporting evidence, conflicting evidence, and whether evidence is
insufficient, giving each equal weight based on the image.

The image and the [Known Perception Ref] supplied in this round jointly form
the basis for judgment. The known Perception Ref consists of visual facts
confirmed to be correct; treat its content directly as correct premises. There
is no need to reverify its truth against the image, and you must not do so. If
the image's surface appearance disagrees with the known Perception Ref, follow
the known Perception Ref. The candidate text itself is not visual evidence;
merely repeating or paraphrasing it does not establish that it is correct.
Evaluate only the current image, question, candidate answer, and known
Perception Ref; do not guess, fill in, or borrow from other candidate answers.

The known Perception Ref covers only the target, reference objects, visual
attributes, and visible locations. It is not the original ground-truth answer
or the final label. It must not supply statements the candidate answer does
not express, and the final label must not be presumed merely because a Ref is
provided.

This is a single-ordinary-option analysis after a multiple-select run, used
for downstream error attribution. Apply the same selection criteria as
multiple-select, with the supplied Ref as additional trusted evidence. Do not
infer or defend the earlier selection, compare options to choose a single best
answer, or infer the model's internal error cause from this response alone.
Different reference objects, wording, or valid paths may yield equally valid
answers. The "None of the above options is correct" option requires all
ordinary options and is outside this single-option task.
```

User Prompt

```
[Single-option attribution task]
candidate_id:
{candidate_id}

Question:
{question}

Candidate answer:
{candidate_answer}

Analyze this ordinary option independently and decide whether it should be
selected under the multiple-select criteria after receiving the supplied Ref.
Keep the binary output labels: correct means select; incorrect means do not
select. An incorrect label due to insufficient evidence does not establish
that the candidate's claims are false.

[Known information type for this round]
Perception Ref (perception_ref)

[Known Perception Ref]
{known_ref}

[Rules for using the known information]
1. The Perception Ref above consists of visual facts confirmed to be correct.
Use it directly; there is no need to reverify its accuracy, and you must not
do so.
2. Continue to judge from the image only those facts not covered by the
Perception Ref.
3. You may cite the known Perception Ref directly in evidence; do not mark it
not_verifiable merely because it is not an image observation.
4. The known Perception Ref provides only the target, reference objects,
visual attributes, and visible locations; it is not the original ground-truth
answer or the final label.
5. Do not use the known Perception Ref to supply spatial relations the
candidate answer does not express, and do not presume correct or incorrect
merely because you have received a Ref.

[Decision order]
1. First directly adopt the facts or derivations in the known Perception Ref,
then identify from the image any targets, reference objects, and relevant
visual facts not covered by it; do not reverify the known Perception Ref.
2. Then verify, one by one, the targets, attributes, reference objects, and
spatial relations explicitly expressed in the candidate. Do not supply,
modify, or correct content that the candidate does not express.
3. Verify the target, attributes, reference objects, and spatial relations,
examining supporting and conflicting evidence using the image and the known
information. Select an option when all its key descriptions are supported and
it can answer the original question.
4. For location questions, a description should effectively narrow down the
target's location; it need not pinpoint a unique location, and neither answer
length nor whether every detail is included is a criterion. Do not exclude an
option merely because its wording is unfamiliar, its reference objects differ,
or it requires further spatial reasoning.
5. Do not select an option if it contains a key contradiction, does not
meaningfully locate the target, or still lacks evidence supporting a key
description after verification. Absence of a contradiction is not sufficient
for selection.

[Evidence requirements]
1. evidence may directly cite facts or derivations in the known Perception
Ref; independent verification from the image is required only for facts not
covered by the known Perception Ref. Merely repeating or paraphrasing the
candidate answer cannot serve as supporting evidence.
```

User Prompt (continued)

```
2. For each key statement in the candidate, examine supporting evidence,
conflicting evidence, and whether evidence is insufficient; do not look for
only one kind.
3. claims must cover every key statement in the candidate that affects
correctness. Do not omit attributes, reference objects, directions, depth,
contact, or strong qualifiers that could make the answer incorrect.
4. When a fact is not covered by the known Perception Ref and image evidence
is insufficient, use not_verifiable; do not fabricate supporting or
contradicting evidence. A key statement that remains not_verifiable prevents
selection, but must not be marked contradicted merely because it lacks
support.
5. Verify additional relations and qualifiers beyond the Ref rather than
extending known facts without justification. The candidate need not repeat
every Ref detail; wording similarity to the Ref is not a selection criterion.
In reason, distinguish a contradiction, failure to answer or meaningfully
locate, and insufficient evidence.

[Coordinate-system rules]
1. First follow any perspective and coordinate requirements explicitly
specified by the question, then determine the coordinate system used by the
current option. A Ref provides evidence within these requirements; it does not
replace the question's task.
2. Subject to the question's constraints, use image coordinates when the
option explicitly specifies image or picture directions.
3. When the candidate does not explicitly specify a coordinate system, use the
reference object's own coordinate system if its intrinsic orientation can be
established; otherwise, prioritize clear scene structure. Use natural
image-based relations only when neither the reference object's own coordinate
system nor clear scene structure applies.
4. If the known Perception Ref does not specify it and the reference object's
orientation or coordinate system cannot be reliably determined, do not force
confirmation of directional relations that depend on it; mark the coordinate
frame as unresolved.
5. Judge front/back and depth relations using orientation, occlusion,
perspective, contact, and scene structure together, not solely from
two-dimensional vertical positions.
6. Relations with strong qualifiers such as "directly," "entirely," or
"tightly against" require more direct evidence from the image or the trusted
Ref than ordinary relations.

[Label mapping]
- All key descriptions are supported and the candidate answers the original
question, meaningfully narrowing down the target's location for a location
question -> correct (select).
- Any key contradiction, failure to answer the question or meaningfully locate
the target, or a key description that remains not_verifiable -> incorrect (do
not select).
- An unresolved coordinate frame prevents selection when a key directional
claim depends on that frame. It is not itself a contradiction; a frame
irrelevant to the candidate's claims is not_applicable.
- Failure to confirm a claim does not establish that it is false. Never use
these binary labels alone to select "None of the above options is correct":
that option requires verification of all ordinary options and cannot be
selected merely because of uncertainty.

Return only valid JSON, without Markdown, code fences, or additional text:
{
  "candidate_id": "Return candidate_id unchanged",
  "relation_analysis": {
    "independent_observation": "Describe relevant image observations and
    explicitly identify any trusted Ref premises used",
    "claims": [
      {
        "claim": "One key statement in the candidate",
        "verdict": "supported, contradicted, or not_verifiable",
```

User Prompt (continued)

```
        "evidence": "Supporting or conflicting image evidence or an explicitly
        cited trusted Ref premise; state when evidence is insufficient"
      }
    ],
    "coordinate_frame": {
      "selected_frame": "object, scene, image, unresolved, or not_applicable",
      "basis": "Question constraints and image evidence or trusted Ref
      premises for this frame; explain any unresolved dependency"
    }
  },
  "label": "correct or incorrect",
  "reason": "Explain why the option should or should not be selected,
  including whether it answers the question and meaningfully locates the
  target; distinguish contradiction from insufficient evidence and cite the
  determining evidence"
}
```

G.2.5.2 Reasoning Ref (reasoning_ref).

System Prompt

```
[Prefix]
You are a neutral visual judge for a spatial-location understanding benchmark.
The candidate answer is a statement to be verified; do not presume it correct
or incorrect before verification is complete. Your task is to examine
supporting evidence, conflicting evidence, and whether evidence is
insufficient, giving each equal weight based on the image.

The image and the [Known Perception and Reasoning Refs] supplied in this round
jointly form the basis for judgment. The known Perception and Reasoning Refs
consist of visual facts, spatial relations, and derivations; treat their
content directly as correct premises. There is no need to reverify their truth
against the image, and you must not do so. If the image's surface appearance
disagrees with the known Perception and Reasoning Refs, follow the known
Perception and Reasoning Refs. The candidate text itself is not visual
evidence; merely repeating or paraphrasing it does not establish that it is
correct. Evaluate only the current image, question, candidate answer, and
known Perception and Reasoning Refs; do not guess, fill in, or borrow from
other candidate answers.

The known Perception and Reasoning Refs cover the target, reference objects,
visual attributes, visible locations, and how spatial relations follow from
these facts. They are not the original ground-truth answer or the final label.
They must not supply statements the candidate answer does not express, and the
final label must not be presumed merely because a Ref is provided.

This is a single-ordinary-option analysis after a multiple-select run, used
for downstream error attribution. Apply the same selection criteria as
multiple-select, with the supplied Ref as additional trusted evidence. Do not
infer or defend the earlier selection, compare options to choose a single best
answer, or infer the model's internal error cause from this response alone.
Different reference objects, wording, or valid paths may yield equally valid
answers. The "None of the above options is correct" option requires all
ordinary options and is outside this single-option task.
```

User Prompt

```
[Single-option attribution task]
candidate_id:
{candidate_id}

Question:
{question}

Candidate answer:
{candidate_answer}

Analyze this ordinary option independently and decide whether it should be
selected under the multiple-select criteria after receiving the supplied Ref.
Keep the binary output labels: correct means select; incorrect means do not
select. An incorrect label due to insufficient evidence does not establish
that the candidate's claims are false.

[Known information type for this round]
Perception and Reasoning Ref (perception_plus_reasoning_ref)

[Known Perception and Reasoning Refs]
[Perceptual facts F]
{perception_input}

[Relational reasoning R]
{reasoning_input}

[Rules for using the known information]
1. The Perception and Reasoning Refs above provide visual facts, spatial
relations, and derivations to use as correct premises. Use them directly;
there is no need to reverify their accuracy, and you must not do so.
2. Continue to judge from the image only those facts not covered by the
Perception and Reasoning Refs.
3. You may cite the known Perception and Reasoning Refs directly in evidence;
do not mark them not_verifiable merely because they are not image
observations.
4. The known Perception and Reasoning Refs provide the target, reference
objects, visual attributes, visible locations, and how spatial relations
follow from these facts; they are not the original ground-truth answer or the
final label.
5. Do not use the known Perception and Reasoning Refs to supply spatial
relations the candidate answer does not express, and do not presume correct or
incorrect merely because you have received a Ref.

[Decision order]
1. First directly adopt the facts or derivations in the known Perception and
Reasoning Refs, then identify from the image any targets, reference objects,
and relevant visual facts not covered by them; do not reverify the known
Perception and Reasoning Refs.
2. Then verify, one by one, the targets, attributes, reference objects, and
spatial relations explicitly expressed in the candidate. Do not supply,
modify, or correct content that the candidate does not express.
3. Verify the target, attributes, reference objects, and spatial relations,
examining supporting and conflicting evidence using the image and the known
information. Select an option when all its key descriptions are supported and
it can answer the original question.
4. For location questions, a description should effectively narrow down the
target's location; it need not pinpoint a unique location, and neither answer
length nor whether every detail is included is a criterion. Do not exclude an
option merely because its wording is unfamiliar, its reference objects differ,
or it requires further spatial reasoning.
5. Do not select an option if it contains a key contradiction, does not
meaningfully locate the target, or still lacks evidence supporting a key
description after verification. Absence of a contradiction is not sufficient
for selection.

[Evidence requirements]
```

User Prompt (continued)

```
1. evidence may directly cite facts or derivations in the known Perception and
Reasoning Refs; independent verification from the image is required only for
facts not covered by the known Perception and Reasoning Refs. Merely repeating
or paraphrasing the candidate answer cannot serve as supporting evidence.
2. For each key statement in the candidate, examine supporting evidence,
conflicting evidence, and whether evidence is insufficient; do not look for
only one kind.
3. claims must cover every key statement in the candidate that affects
correctness. Do not omit attributes, reference objects, directions, depth,
contact, or strong qualifiers that could make the answer incorrect.
4. When a fact is not covered by the known Perception and Reasoning Refs and
image evidence is insufficient, use not_verifiable; do not fabricate
supporting or contradicting evidence. A key statement that remains
not_verifiable prevents selection, but must not be marked contradicted merely
because it lacks support.
5. Verify additional relations and qualifiers beyond the Ref rather than
extending known facts without justification. The candidate need not repeat
every Ref detail; wording similarity to the Ref is not a selection criterion.
In reason, distinguish a contradiction, failure to answer or meaningfully
locate, and insufficient evidence.

[Coordinate-system rules]
1. First follow any perspective and coordinate requirements explicitly
specified by the question, then determine the coordinate system used by the
current option. A Ref provides evidence within these requirements; it does not
replace the question's task.
2. Subject to the question's constraints, use image coordinates when the
option explicitly specifies image or picture directions.
3. When the candidate does not explicitly specify a coordinate system, use the
reference object's own coordinate system if its intrinsic orientation can be
established; otherwise, prioritize clear scene structure. Use natural
image-based relations only when neither the reference object's own coordinate
system nor clear scene structure applies.
4. If the known Perception and Reasoning Refs do not specify it and the
reference object's orientation or coordinate system cannot be reliably
determined, do not force confirmation of directional relations that depend on
it; mark the coordinate frame as unresolved.
5. Judge front/back and depth relations using orientation, occlusion,
perspective, contact, and scene structure together, not solely from
two-dimensional vertical positions.
6. Relations with strong qualifiers such as "directly," "entirely," or
"tightly against" require more direct evidence from the image or the trusted
Ref than ordinary relations.

[Label mapping]
- All key descriptions are supported and the candidate answers the original
question, meaningfully narrowing down the target's location for a location
question -> correct (select).
- Any key contradiction, failure to answer the question or meaningfully locate
the target, or a key description that remains not_verifiable -> incorrect (do
not select).
- An unresolved coordinate frame prevents selection when a key directional
claim depends on that frame. It is not itself a contradiction; a frame
irrelevant to the candidate's claims is not_applicable.
- Failure to confirm a claim does not establish that it is false. Never use
these binary labels alone to select "None of the above options is correct":
that option requires verification of all ordinary options and cannot be
selected merely because of uncertainty.

Return only valid JSON, without Markdown, code fences, or additional text:
{
  "candidate_id": "Return candidate_id unchanged",
  "relation_analysis": {
    "independent_observation": "Describe relevant image observations and
    explicitly identify any trusted Ref premises used",
```

User Prompt (continued)

```
    "claims": [
      {
        "claim": "One key statement in the candidate",
        "verdict": "supported, contradicted, or not_verifiable",
        "evidence": "Supporting or conflicting image evidence or an explicitly
        cited trusted Ref premise; state when evidence is insufficient"
      }
    ],
    "coordinate_frame": {
      "selected_frame": "object, scene, image, unresolved, or not_applicable",
      "basis": "Question constraints and image evidence or trusted Ref
      premises for this frame; explain any unresolved dependency"
    }
  },
  "label": "correct or incorrect",
  "reason": "Explain why the option should or should not be selected,
  including whether it answers the question and meaningfully locates the
  target; distinguish contradiction from insufficient evidence and cite the
  determining evidence"
}
```