PERCEPTVLM: FROM SEEING DETAILS TO CONNECTING EVIDENCE FOR GENERALIZABLE MULTIMODAL PERCEPTION

Anonymous authors
Paper under double-blind review

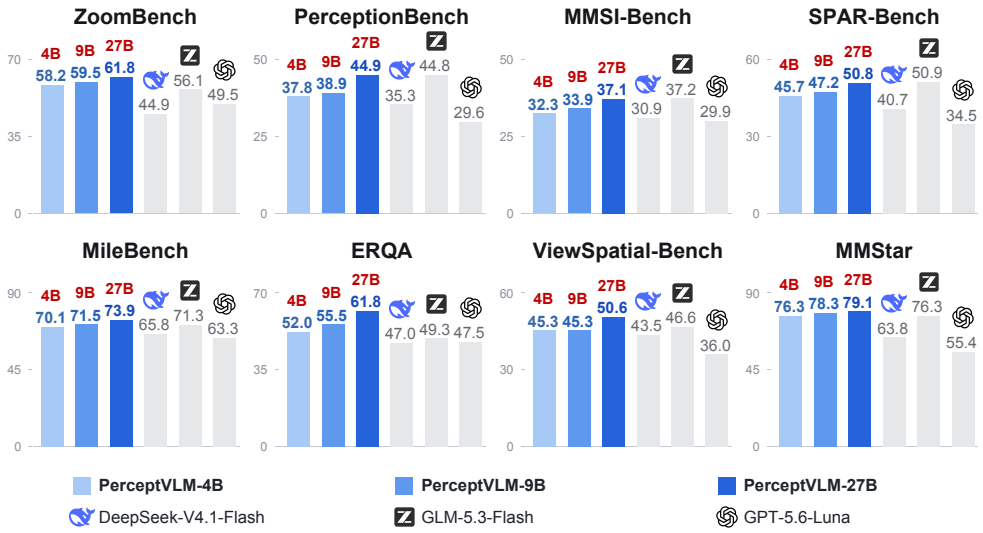


Figure 1: Performance comparison of PerceptVLM and other advanced models.

ABSTRACT

Multimodal perception forms the foundation of multimodal intelligence. In particular, spatial perception and multimodal long-context understanding are two pivotal tasks that underpin embodied AI and multimodal agents. However, synthesizing targeted, high-quality training data for these tasks remains difficult and costly. Our key insight is that these tasks share an underlying capability, which we call connecting evidence: identifying relevant visual facts, binding them to the correct entities, recovering their relations, and composing them into an answer. Crucially, targeted training data are far easier to synthesize for this capability than for these tasks themselves. To strengthen this capability, we propose EVIDENCECONNECT, an easy-to-implement data synthesis framework that uses zoomed views of related regions within a single image to generate questions requiring comparison, aggregation, or chaining of visual evidence. To ensure quality, a reference answer is retained only when independent VLMs agree on it, and region-ablation tests discard questions whose answer survives the removal of selected regions. We then train base VLMs on the resulting data with RLVR, obtaining PerceptVLM. Although the synthesized data are not tailored to any downstream benchmark, PerceptVLM improves fine-grained and spatial perception, with further gains on embodied spatial reasoning, multimodal long-context understanding, multimodal agent tasks, and general visual reasoning (shown in Figure 1). These results establish EVIDENCECONNECT as an effective framework for supervising evidence connection and suggest that strengthening this shared capability leads to more generalizable multimodal perception.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Multimodal perception is the foundation of multimodal intelligence: before a multimodal model can reason or act, it must determine what is present in a scene. Recent progress on this front, such as multimodal Chain-of-Thought (Zhang et al., 2023; Xu et al., 2024; Li & Ma, 2026) and Thinking-with-Images (Zheng et al., 2025; Zhang et al., 2025b), has improved the ability of open-source vision-language models (VLMs) to perceive fine-grained visual content and use it in reasoning.

However, these advances do not reliably extend to tasks in which critical visual cues are dispersed across objects, relationships, frames, or dialogue turns. Two such tasks, spatial perception and multimodal long-context understanding, are particularly important because they underpin embodied AI and multimodal agents, respectively. Spatial perception requires a model to recover how entities are arranged relative to one another and how several spatial relations compose (Gemini Robotics Team, 2025; Li et al., 2025a). Embodied AI relies on it to ground navigation, manipulation, and viewpoint-dependent decisions in object locations and spatial relations (Brohan et al., 2023; Kim et al., 2025; Yang et al., 2026). Multimodal long-context understanding requires a model to locate evidence across images and dialogue history, preserve entity identity, and bring that evidence to bear on the current question (Wu et al., 2025; Liu et al., 2024; Wang et al., 2025d; Song et al., 2024). Multimodal agents rely on it to retrieve relevant past observations, resolve references across interactions, and integrate earlier evidence into current decisions (He et al., 2024; Su et al., 2026; Li et al., 2025c).

Strengthening these capabilities, however, is constrained by the difficulty of building high-quality training data. For spatial perception, existing efforts mainly scale up region-grounded relational data (Chen et al., 2024a; Zhang et al., 2026b), but reliable labels require precise localization and relational grounding across multiple objects. Such labels are expensive to annotate by hand, while automated generators often hallucinate relations. For long-context understanding, existing multi-image and multi-turn instruction data (Jiang et al., 2024; Liu et al., 2024) concentrate on a few formats and task types, because diverse questions that remain answerable over long sequences are hard to construct automatically. Data for embodied spatial reasoning and multimodal agents are even harder to synthesize, since they additionally involve physical or interactive environments and multi-step trajectories. Consequently, neither these capabilities nor the applications built on them have a scalable and high-quality source of clean, verifiable supervision.

Our key insight is that these tasks build on a shared atomic capability of VLMs, which we call *connecting evidence*: identifying task-relevant visual facts, binding each fact to the correct entity, recovering the relations among them, and integrating the evidence into an answer. These tasks instantiate this capability differently: spatial perception connects objects within a scene through spatial relations, whereas long-context understanding, on which multimodal agents rely, connects facts across images and dialogue turns through entity identity and reference. Despite these differences in where the evidence resides and how it is linked, both tasks rely on the same underlying steps, so strengthening these steps can benefit both. Moreover, since the first step, identifying visual facts, is itself fine-grained perception, fine-grained perception benefits as well.

Crucially, training data that target this capability are far easier to construct and verify than data for the tasks themselves. To turn this insight into trainable supervision, we propose EVIDENCECONNECT, an easy-to-implement, fully automated framework that synthesizes visual questions requiring the model to connect evidence across regions. Specifically, we select a group of related regions within a single image and synthesize a compositional question over them from *privileged* zoomed views, which make small details visible but are used only during data construction. Depending on the scene, we synthesize questions to compare, aggregate, or chain evidence across the regions. Answering such questions thus exercises the same steps as the target tasks: relating regions spatially, including from a given viewpoint, parallels spatial perception and embodied spatial reasoning, whereas locating scattered regions, binding their facts to the right references, and integrating them parallels long-context understanding and multimodal agents. To ensure data quality, we keep a question only if VLMs from different families agree on its answer, and we use region-ablation tests to discard questions whose answer survives the removal of selected regions. Finally, each verified question is converted into multiple-choice form for rule-based rewards and paired with the full image, where colored outlines typically mark the selected regions to avoid ambiguous references, so the model must still perceive and connect the evidence by itself.

To evaluate whether this supervision generalizes beyond the synthesized questions, we train Qwen3.5-4B, Qwen3.5-9B, and Qwen3.6-27B (Qwen Team, 2026a;b) with reinforcement learning with verifiable rewards (RLVR) (Lambert et al., 2024; DeepSeek-AI, 2025) using only our synthesized data, without supervised fine-tuning or benchmark-specific question formats. The resulting models, PerceptVLM-4B/9B/27B, achieve substantial improvements in fine-grained and spatial perception as well as multimodal long-context understanding, with gains that further extend to embodied spatial reasoning, multimodal agent tasks, and general visual reasoning. For example, PerceptVLM-4B improves over its base model from 37.43 to 45.26 (+7.83) on ViewSpatial (spatial reasoning) and from 60.08 to 70.08 (+10.00) on MileBench (long-context understanding). More broadly, its average over all 26 benchmarks increases from 48.10 to 53.56, and the corresponding averages improve from 50.30 to 55.30 for PerceptVLM-9B and from 54.32 to 59.23 for PerceptVLM-27B. Notably, although PerceptVLM is trained only on single images, it also improves on multi-image long-context benchmarks and multimodal agent tasks. These results demonstrate that connecting evidence is indeed a generalizable training target for multimodal perception.

To summarize, our key contributions are the following:

- We identify connecting evidence as an important atomic capability of VLMs, underlying fine-grained perception, spatial perception, and multimodal long-context understanding..
- We develop EVIDENCECONNECT, a structured and audited synthesis pipeline for single-image, multi-region compositional questions, with privileged region views and independent checks for answer correctness and evidence dependence.
- Experiments demonstrate that RLVR on the resulting synthetic data improves a broad range of perception benchmarks and further generalizes to many real-world tasks, including embodied spatial reasoning, multimodal agent tasks, and general visual reasoning.

2 RELATED WORK

Spatial perception. Spatial perception underpins embodied tasks such as robotic manipulation and embodied question answering (Brohan et al., 2023; Kim et al., 2025; Gemini Robotics Team, 2025). Existing approaches use large-scale spatial supervision (Chen et al., 2024a), intermediate visual grounding (Zhang et al., 2026b), and progressive training (Li et al., 2026). Benchmarks and associated training resources examine perspective-taking, multi-image inference, and multi-hop composition (Li et al., 2025a; Yang et al., 2026). MultihopSpatial additionally studies grounded spatial training and its transfer to embodied tasks (Lee et al., 2026). Euclid's Gift (Lian et al., 2026) demonstrates transfer from geometric surrogate tasks. Similarly, we train on a surrogate task and evaluate its transfer, but our task is built from natural images and targets the evidence connection shared by spatial perception and long-context understanding, rather than spatial skills alone.

Long-context multimodal understanding. Multimodal long-context understanding requires locating, tracking, and integrating evidence across long sequences of images and dialogue turns. Existing efforts support these abilities through multi-image and multi-turn instruction tuning (Jiang et al., 2024; Huang et al., 2023; Liu et al., 2024) and dialogue-fragment retrieval (Bi et al., 2026), while benchmarks assess long-context integration and memory (Song et al., 2024; Wang et al., 2025d; Ren et al., 2026). Controlled studies nevertheless identify persistent difficulties in aggregating distributed visual information (Das et al., 2026). We target evidence connection as a shared component of these tasks, constructing single-image questions that require identifying, binding, and composing facts across regions. This setting makes the selected evidence easier to localize and audit, allowing us to investigate whether such supervision transfers to multimodal long-context understanding without training on long image sequences or dialogue histories.

Multimodal synthetic data. Early resources provide annotated scene graphs and programmatic questions (Krishna et al., 2017; Hudson & Manning, 2019). Later methods automate instruction generation (Zhao et al., 2024; Zhang et al., 2025a), evolution (Luo et al., 2025), and synthesis for unsupervised post-training (Wei et al., 2025). Zooming without Zooming (Wei et al., 2026a) transfers single-region supervision to full images with cross-model answer checks. CRIT (Sung et al., 2026) and FM2DS (Abaskohi et al., 2024) synthesize cross-modal questions and filter unimodal shortcuts, while MM-CondChain (Shen et al., 2026) verifies conditional programs and ReasonVQA (Tran et al., 2025) composes external knowledge paths. HopChain (Wang et al., 2026) targets long-CoT robustness through instance-dependent chains with human-verified answers. Unlike these pipelines,

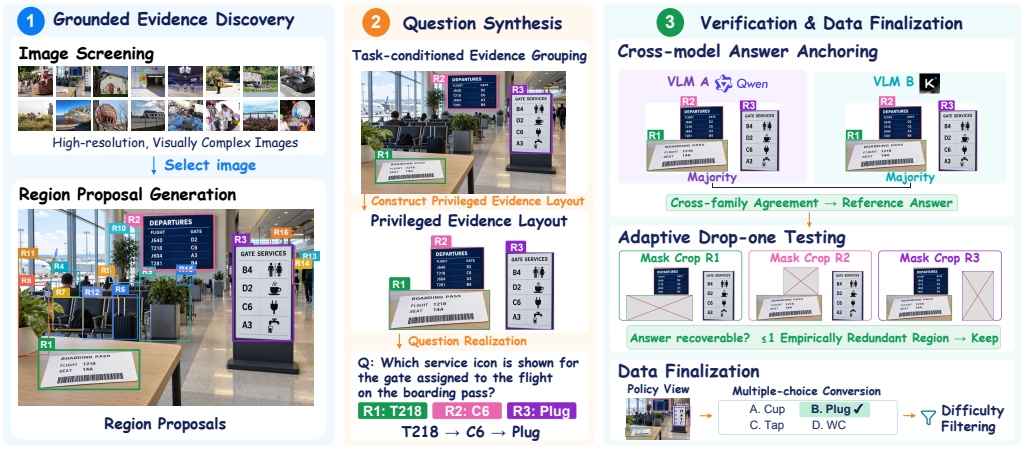


Figure 2: Overview of our multi-region question synthesis pipeline. We select related regions from a single image and construct a privileged crop view to generate compositional questions. The illustrative example connects a flight number on a boarding pass to its departure gate and then to a service icon (T218 → C6 → plug). Cross-model answer anchoring and drop-one region tests assess answer reliability and evidence dependence, followed by multiple-choice conversion and difficulty filtering. During training, the policy receives the full image with region outlines, while privileged crops are used only for data construction.

EVIDENCECONNECT builds each question on a coherent group of regions rather than a single region or a sequential chain, so its questions cover aggregation and closed-set comparison in addition to dependency chains. Moreover, it automatically checks both the answer and its dependence on the selected regions, through cross-model agreement and region ablation, without human annotation.

3 BOOSTING MULTIMODAL PERCEPTION BY LEARNING TO CONNECT VISUAL EVIDENCE

As discussed in Section 1, fine-grained perception, spatial perception, and multimodal long-context understanding share a core capability that we call connecting evidence. This motivates training data whose questions structurally require this capability and end in concise answers suitable for verifiable rewards. To construct this supervision, we propose EVIDENCECONNECT, a data synthesis framework centered on coherent groups of image regions. Within this framework, high-resolution privileged crops render local details visible for question synthesis, while cross-model answer anchoring and region ablation check label reliability and evidence dependence. The overview of the data synthesis pipeline is shown in Figure 2.

3.1 MULTI-REGION QUESTION SYNTHESIS

Target question structure. We first define the target question, whose answer depends on connecting one coherent group of regions in an image. Formally, from a source image I, we obtain localized region proposals R_I and select one evidence group $G = \{r_{i_1}, \ldots, r_{i_k}\} \subseteq R_I, 3 \leq k \leq 7$. Restricting the evidence to a single image keeps training inexpensive, while several related regions of one image already require a model to locate, bind, and compose dispersed evidence. Importantly, the regions are selected to jointly support one question q, which assigns each of them a role, i.e., the fact it contributes: a value to be combined, a member of a closed candidate set, or a clue on which a later step depends.

Corresponding to these three kinds of roles, q connects the regions through one of three forms. *Aggregation* combines counts or other numeric facts from the regions, possibly through a short calculation. *Comparison and matching* treats the regions as the complete candidate set, so that an answer such as an outlier, an extremum, or a pairing can be identified only by considering all of them. *Dependency chains and compositions* build the answer in several steps, e.g., by using a clue

to decide what to read next, fixing a viewpoint for interpreting the rest, or combining predicates from different regions. In all three forms, the answer is determined only when the assigned roles are taken together. These forms also tie our questions to the target tasks: spatial relations, including those in an agent's or object's own frame, parallel spatial perception and embodied spatial reasoning, while binding facts to regions referenced by outline colors and integrating them parallels how long-context understanding and multimodal agents handle facts tied to image or turn references. To synthesize such questions, we first locate candidate regions, then group those that support one of the three forms, and finally compose questions over each selected group.

Grounded evidence discovery. Since multi-region questions need detail-rich scenes, we start from high-resolution, visually complex images. For each image, a strong VLM judges whether it can support such questions and, if so, is prompted to propose concrete, clearly localizable regions, each with a bounding box and a short description, prioritizing regions that can be related to one another over isolated salient ones. We then discard proposals that are too small, too large, or heavily overlapping. As these model-generated descriptions may be wrong, they only guide grouping and question writing, whereas reference answers are later derived from the image itself.

Task-conditioned evidence grouping. An arbitrary collection of these proposals, however, does not necessarily support a valid compositional question. We therefore prompt a strong model with the image caption and region descriptions to propose several groups per image, each labeled with a task family (a specific question type such as anomaly detection) and a necessity mode (how removing a region should affect the answer). We then discard invalid or duplicated groups and closed sets of incomparable objects, and keep groups with distinct task families first. The task family only guides question writing, whereas each region's concrete role is fixed when a question is written.

Privileged evidence layout. Since fine-grained details are difficult to inspect at full-image scale, we construct for each group a privileged evidence layout V_G, which is accessible only during data construction. Following the layout-preserving design of Liu et al. (2025), V_G arranges native-resolution crops of the selected regions, with small context margins, on a compact canvas that removes irrelevant background while preserving their relative horizontal and vertical order. In addition, each region is outlined in a distinct color that also marks it in the full image, allowing questions to refer to regions by color unambiguously. Finally, V_G serves as the sole visual input to the question generator, which confines its questions to evidence within the group, and it is later reused for verification (Section 3.2).

Question realization. The question generator, a strong VLM, then receives V_G, the region descriptions, and a few reasoning blueprints, i.e., templates of reasoning programs, and composes several candidate questions, each with a concise answer: a single value (e.g., a number, a short phrase, or the outline color of a selected region), a set, or an ordered pair. To make its use of evidence checkable, the generator also states each region's contribution and the answer expected after removing it, together with a reasoning certificate, i.e., a short sequence of steps that links the regions to the answer. Deterministic rules then reject candidates whose certificate leaves a region unused or contains steps that do not lead to the answer, whose own predicted answer survives the removal of a region, or that use subjective wording, leak the answer through an outline color, or contain inconsistent arithmetic. Implementation details and prompt templates appear in Appendices A and B.

3.2 Two-stage Verification

The checks above concern only the form of a candidate. A compositional question, however, is useful supervision only if its reference answer is reliable and the intended evidence matters. We assess these properties separately, since a plausible reasoning structure establishes neither.

Cross-model answer anchoring. To obtain a more reliable reference answer, two VLMs from different model families answer q from V_G several times, without seeing the region descriptions or the proposed answer. A candidate is kept only if a clear majority of each model's responses supports the same answer and this answer becomes the reference $a^\star$.

Adaptive drop-one testing. Agreement on $a^\star$, however, does not show that the answer actually depends on the selected evidence. Since every question is constructed to require all of its regions (Section 3.1), an answer that survives the removal of a region indicates a defect, such as a decorative region or a shortcut that bypasses the intended reasoning. To detect such defects, we mask each $r_i \in G$ in the privileged layout, $V_G^{-i} = \mathcal{M}_i(V_G)$, where $\mathcal{M}_i$ replaces the crop of r_i with a gray placeholder. A verifier from one anchoring family then answers q from this view, reporting insufficient information if the masked evidence is needed, and either such a report or a changed answer confirms that the question depends on r_i. If instead $a^\star$ is still returned, or the response is unusable, we query the verifier again and mark r_i as *empirically redundant* only if most queries recover $a^\star$. Note that a single redundant region, however, may reflect noise rather than a defect, because part of a masked region can remain visible in an adjacent crop through overlapping context margins, and individual responses also vary. We therefore reject a question only when more than one region is redundant, which also improves synthesis efficiency, since a stricter rule would discard many potentially valid questions merely because one region was judged redundant. The implementation details are shown in Appendix A.

3.3 Data Finalization

From privileged crops to the policy view. Since the policy, i.e., the model under training, never sees V_G, each verified candidate is paired with a policy view V_I, by default the source image with each selected region outlined in its assigned color. To avoid ambiguous references in the full image, these outlines mark which instances a question refers to, without revealing the answer-bearing content that the policy must still perceive and compose. For aggregate questions, whose answers do not depend on the outlines, we also try to replace marker references with natural descriptions, so that the policy must locate the evidence itself. This unmarked version replaces the marked one only if independent audits by both anchoring families, without access to $a^\star$, confirm unique references, an unchanged multi-region task grounded in the intended regions, and the answer $a^\star$.

Multiple-choice conversion and difficulty filtering. To support rule-based rewards in RLVR, we then turn each question into a four-option multiple-choice question, adding three distractors in the answer format of $a^\star$ so that the format does not reveal the answer. Each released example is thus $x = (V_I, q_{\text{rel}}, O)$ with label $y \in \{\text{A}, \text{B}, \text{C}, \text{D}\}$, where q_{rel} is the verified question stem, rewritten without marker references when the unmarked view is used, O the shuffled option set, and y the option of $a^\star$. Finally, since overly easy examples provide little learning signal, a lightweight VLM answers each example several times, and we discard examples that it always answers correctly.

4 Experiments

This section evaluates whether learning to connect evidence within a single image yields improvements that generalize across multimodal perception tasks.

4.1 Experimental Settings

Benchmarks. We evaluate on 26 benchmarks in six domains (Table 1). *(i) Fine-grained perception:* VStar (Wu & Xie, 2024), HR-Bench 4K/8K (Wang et al., 2025c), MME-RealWorld (EN/CN) (Zhang et al., 2025c), ZoomBench (Wei et al., 2026a), Chartography (Garre et al., 2026), and PerceptionBench (Lin et al., 2026). *(ii) Spatial perception:* TreeBench (Wang et al., 2025a), O3-Bench (Li et al., 2025b), CountQA (Tamarapalli et al., 2025), MIMIC (Das et al., 2026), MMSI-Bench (Yang et al., 2026), and SPAR-Bench (Zhang et al., 2026a). *(iii) Multimodal long-context understanding:* MMLongBench (Wang et al., 2025d), MileBench (Song et al., 2024), MemLens (Ren et al., 2026), and CL-Bench-V (Wei et al., 2026b). *(iv) Embodied spatial reasoning:* ERQA (Gemini Robotics Team, 2025), ViewSpatial (Li et al., 2025a), and MindCube-Tiny (Wang et al., 2025b). *(v) Multimodal agents:* AgentVista (Su et al., 2026), together with ScreenSpot-Pro (Li et al., 2025c) and UI-Vision (Nayak et al., 2025) for GUI grounding. *(vi) General visual reasoning:* MMStar (Chen et al., 2024b) and BabyVision (Chen et al., 2026). We report accuracy (%) on all benchmarks.

Training. We train Qwen3.5-4B/9B (Qwen Team, 2026a) and Qwen3.6-27B (Qwen Team, 2026b) in non-thinking mode with GRPO (Shao et al., 2024) on the 20.4K examples synthesized by Ev-

Table 1: Main results on 26 benchmarks (accuracy, %). Domain averages and Overall are unweighted means within each domain and over all 26 benchmarks, respectively.

Domain	Benchmark	Base models			PerceptVLM			Advanced models		
		4B	9B	27B	4B	9B	27B	GPT	GLM	DeepSeek
Fine-grained perception	VStar	82.20	83.77	87.96	89.00	90.58	91.62	62.83	79.58	70.68
	HR-4K	82.37	84.88	84.25	88.62	89.75	89.88	66.38	82.50	71.77
	HR-8K	78.37	82.50	80.00	83.63	87.62	88.00	53.50	76.13	62.72
	MME-en	70.60	71.00	72.62	75.48	75.33	77.41	58.50	67.88	50.75
	MME-cn	69.33	69.53	68.68	73.89	73.45	74.97	56.90	68.31	50.04
	ZoomBench	50.41	49.57	53.61	58.22	59.53	61.78	49.53	56.09	44.85
	Chartography	5.00	7.00	8.00	14.00	13.00	12.00	9.09	9.00	8.00
	Perception-Bench	35.53	38.70	41.57	37.80	38.89	44.93	29.57	44.78	35.28
	Avg.	59.23	60.87	62.09	**65.08**	**66.02**	**67.57**	48.29	60.53	49.26
Spatial perception	TreeBench	50.37	46.67	49.63	51.36	51.60	56.05	47.28	57.53	49.26
	O3-Bench	53.91	59.13	65.80	63.77	67.83	72.46	38.55	60.00	42.90
	CountQA	36.91	37.63	46.73	47.19	48.43	54.32	21.34	49.35	47.51
	MIMIC	55.01	59.39	55.95	66.10	64.98	65.16	57.43	73.46	68.55
	MMSI-Bench	30.70	32.70	33.50	32.30	33.90	37.10	29.90	37.20	30.85
	SPAR-Bench	41.47	43.53	46.68	45.69	47.23	50.78	34.53	50.94	40.66
	Avg.	44.73	46.51	49.72	**51.07**	**52.33**	**55.98**	38.17	54.75	46.62
Long-context understanding	MMLongBench*	63.50	69.89	73.41	68.34	73.45	76.73	–	–	–
	MileBench	60.08	68.96	71.97	70.08	71.46	73.94	63.34	71.33	65.79
	MemLens	40.62	45.82	52.22	42.59	47.34	55.64	–	–	–
	CL-Bench-V	15.36	17.26	24.18	22.85	23.25	27.30	–	–	–
	Avg.	44.89	50.48	55.45	**50.97**	**53.88**	**58.40**	–	–	–
Embodied spatial reasoning	ERQA	45.75	45.75	51.25	52.00	55.50	61.75	47.50	49.25	47.00
	ViewSpatial-Bench	37.43	42.65	48.49	45.26	45.27	50.63	35.98	46.57	43.51
	MindCube-Tiny	47.62	40.76	50.76	53.05	52.57	54.76	42.57	65.71	36.00
	Avg.	43.60	43.05	50.17	**50.10**	**51.11**	**55.71**	42.02	53.84	42.17
Multimodal agent	AgentVista	12.44	8.61	12.92	17.22	14.83	18.18	–	19.62	–
	ScreenSpot-Pro	60.09	63.88	70.21	61.73	66.98	70.59	–	–	–
	UI-Vision	35.83	42.48	66.05	37.67	45.48	66.66	–	–	–
	Avg.	36.12	38.32	49.73	**38.87**	**42.43**	**51.81**	–	–	–
General visual reasoning	MMStar	74.13	77.27	75.93	76.27	78.27	79.07	55.41	76.33	63.77
	BabyVision	15.46	18.56	19.85	18.56	21.39	28.35	13.14	29.12	20.00
	Avg.	44.80	47.92	47.89	**47.42**	**49.83**	**53.71**	34.28	52.73	41.89
Overall		48.10	50.30	54.32	**53.56**	**55.30**	**59.23**	–	–	–

Notes. Base models are Qwen3.5-4B/9B and Qwen3.6-27B. GPT: GPT-5.6-Luna; GLM: GLM-5.3-Flash. DeepSeek: DeepSeek-V4.1-Flash. All models are evaluated in low-thinking or non-thinking reasoning modes. *: extracted MMLongBench visual-core subset, including V-RAG, Image MM-NIAH, and VH-Multi (25 non-Counting tasks in total).

IDENCECONNECT (Appendix C), without supervised fine-tuning. All three scales share the same data and training configuration (Appendix D), and we refer to the resulting models as PerceptVLM-4B/9B/27B.

Baselines. We compare with two groups of baselines. As references for absolute performance, Table 1 includes three advanced models: GPT-5.6-Luna, GLM-5.3-Flash, and DeepSeek-V4.1-Flash. To compare EVIDENCECONNECT with existing synthetic training data, we further train Qwen3.5-4B with the same GRPO recipe on the datasets of ZwZ (Wei et al., 2026a), TVG (Zhang et al., 2026b), F^2RVLM (Bi et al., 2026), CRIT (Sung et al., 2026), and MMHops (Zhang et al., 2026c) (Table 2). For CRIT, whose 84,199 training samples contain 269,467 QA pairs from GQA, ActivityNet Captions, and SPIQA in multi-image, interleaved image–text, and multi-turn formats, we uniformly sample about 30K QA pairs across the three domains and train on them without CoT supervision or additional annotations.

4.2 MAIN RESULTS

Table 1 shows that PerceptVLM improves over its base model on all 26 benchmarks at every scale, raising the overall average by 5.46, 5.00, and 4.91 points at 4B, 9B, and 27B, respectively. In particular, fine-grained and spatial perception improve by 5.2–6.3 points on average, with gains of up to 11.1 points on MIMIC and 10.8 points on CountQA, and embodied spatial reasoning improves by 5.5–8.1 points, including 10.5 points on ERQA at 27B. Qualitative examples are provided in Appendix F.

Table 2: Training Qwen3.5-4B on different synthetic datasets (accuracy, %). Parentheses give the number of training examples.

	Fine-grained perception			Spatial perception			Embodied spatial reasoning			Long-context understanding		
Model	**HR-4K**	**HR-8K**	**Chart**	**O3**	**CountQA**	**MMSI**	**MindCube**	**ERQA**	**ViewSpatial**	**MileBench**	**MemLens**	**Avg.**
Base	82.37	78.37	5.00	53.91	36.91	30.70	47.62	45.75	37.43	60.08	40.62	47.16
ZwZ (73k)	87.00	82.75	7.00	60.87	43.32	31.33	33.81	46.00	43.75	70.21	42.46	49.86
TVG (19.4k)	78.25	73.62	13.00	53.04	33.97	29.50	28.86	46.00	42.16	64.56	37.29	45.48
F2RVLM (37k)	80.13	77.62	10.00	54.20	32.72	31.90	34.38	44.75	40.56	64.58	39.80	46.42
CRIT (30k)	82.25	77.88	6.00	45.22	30.56	29.90	43.43	45.00	43.98	68.13	40.40	46.61
MMHops (31k)	77.75	73.12	3.00	45.80	30.82	30.50	42.10	43.75	40.90	66.29	15.97	42.73
Ours (20.4k)	88.62	83.63	14.00	63.77	47.19	32.30	53.05	52.00	45.26	70.08	42.59	**53.86**

Table 3: Effect of privileged views on Qwen3.5-4B (accuracy, %). Questions are synthesized either directly from the original image or, as in EVIDENCECONNECT, from privileged zoomed views of the selected regions. Both variants are trained on 2.4K examples, rather than the full 20.4K training set used in Table 2.

	Fine-grained perception			Spatial perception			Embodied spatial reasoning			Long-context understanding		
Synthesized from	**HR-4K**	**HR-8K**	**Chart**	**O3**	**CountQA**	**MMSI**	**MindCube**	**ERQA**	**ViewSpatial**	**MileBench**	**MemLens**	**Avg.**
Base	82.37	78.37	5.00	53.91	36.91	30.70	47.62	45.75	37.43	60.08	40.62	47.16
Original image	81.37	79.00	12.00	58.84	35.67	31.60	32.29	50.00	44.29	69.11	40.94	48.65
Privileged views	84.62	81.25	8.00	61.74	37.24	33.40	30.76	52.00	42.47	68.48	41.03	**49.18**

Notably, although PerceptVLM is trained only on single images, its gains carry over to multi-image and interactive settings. Long-context understanding improves at every scale, most markedly at 4B, where the domain average rises by 6.1 points with gains of 10.0 points on MileBench and 7.5 points on CL-Bench-V. Likewise, the multimodal agent benchmark AgentVista improves by 4.8–6.2 points, and GUI grounding also improves consistently. These results support our premise that connecting evidence within an image transfers to tasks whose evidence spans images, dialogue turns, or interaction steps.

PerceptVLM demonstrates competitive performance against frontier models. Specifically, PerceptVLM-27B surpasses GPT-5.6-Luna, GLM-5.3-Flash, and DeepSeek-V4.1-Flash in average accuracy across all evaluated dimensions, including fine-grained perception, spatial perception, embodied spatial reasoning, and general visual reasoning.

4.3 COMPARISON WITH EXISTING TRAINING DATA

Table 2 compares EVIDENCECONNECT with existing synthetic datasets on the same 4B base model. These datasets target single-region fine-grained perception (ZwZ), grounded counting and spatial reasoning (TVG), dialogue-fragment retrieval (F^2RVLM), cross-modal multi-hop QA over interleaved contexts (CRIT), and knowledge-based multi-hop QA (MMHops). Training on TVG, F^2RVLM, CRIT, or MMHops improves MileBench but degrades HR-Bench, CountQA, MindCube, and MemLens, and all four datasets lower the average accuracy below that of the base model. In contrast, EVIDENCECONNECT improves every benchmark over the base model and outperforms these four datasets on all 11 benchmarks, with a comparable amount of data (20.4K examples versus 19.4K–37K).

ZwZ-73K is the strongest baseline. Since it also synthesizes questions from zoomed-in views but grounds each question in a single region, the comparison with ZwZ serves as an approximate ablation of connecting multiple regions. Although ZwZ uses 3.6× more examples (73K versus 20.4K), EVIDENCECONNECT raises the average accuracy from 49.86 to 53.86 and surpasses ZwZ on 10 of the 11 benchmarks, trailing it only on MileBench by 0.13 points. The largest margin appears on MindCube, where ZwZ reduces accuracy by 13.8 points whereas EVIDENCECONNECT improves it by 5.4 points. This contrast is consistent with our premise that embodied spatial reasoning relies on connecting evidence across regions rather than on inspecting a single region.

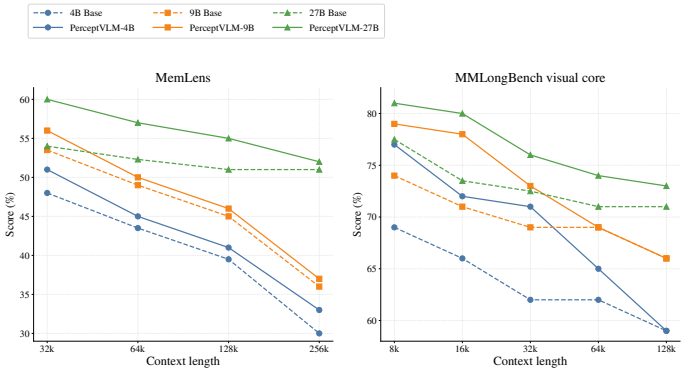


Figure 3: Context-length performance on MemLens (left) and the MMLongBench visual core (right). Dashed and solid lines show the base models and PerceptVLM, respectively.

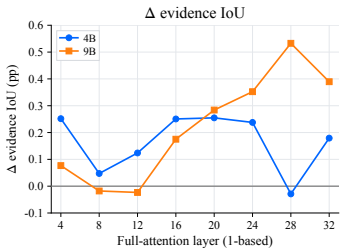


Figure 4: **Attention alignment on TreeBench.** Layer-wise change in mean top-k evidence IoU (PerceptVLM minus Qwen3.5, percentage points).

4.4 Analysis

Ablation study on privileged views. To examine the role of privileged views, we train Qwen3.5-4B on 2.4K examples synthesized from privileged views or original images, with all other settings fixed (Table 3). Privileged views yield higher average accuracy, with gains of 1.6–3.3 points on HR-Bench 4K/8K, O3-Bench, CountQA, and MMSI-Bench. Without them, HR-Bench 4K and CountQA fall below the base model. The variants perform similarly on long-context tasks and show mixed results on embodied benchmarks, suggesting that privileged views strengthen supervision for fine-grained visual evidence.

Context length. Figure 3 breaks down MemLens and the MMLongBench visual core by context length. PerceptVLM outperforms or matches its base model at every length and scale, and on MemLens its gains persist up to 256K tokens. The gains, however, narrow at the longest contexts. At 128K on the MMLongBench visual core, for example, PerceptVLM-4B and PerceptVLM-9B merely match their base models. This trend is consistent with our training data, in which each example contains a single image and thus a short context. Training therefore strengthens evidence connection but not retrieval from very long inputs, which likely becomes the main difficulty at the longest lengths.

Attention to evidence. To examine whether PerceptVLM focuses more on task-relevant visual evidence, we follow Wei et al. (2026a) and evaluate PerceptVLM-4B/9B against their base models on 405 TreeBench (Wang et al., 2025a) examples equipped with annotated evidence bounding boxes. For each full-attention layer, we compute the Intersection over Union (IoU) between the ground-truth evidence tokens E and the $|E|$ visual tokens with the highest attention weights (Appendix E). As shown in Figure 4, PerceptVLM achieves a higher mean IoU across 7 out of 8 layers at the 4B scale and 6 out of 8 layers at the 9B scale. This indicates that the high-attention tokens of PerceptVLM align more closely with the ground-truth evidence in most layers at both scales.

5 Conclusion

We presented EvidenceConnect, a fully automated framework that synthesizes multi-region visual questions and answers, requiring a model to connect evidence within an image. Trained with RLVR on these data alone, PerceptVLM improves over its base models on all 26 benchmarks at 4B, 9B, and 27B. Notably, although training uses only single images, the gains extend beyond fine-grained and spatial perception to embodied spatial reasoning, multimodal long-context understanding, and multimodal agent tasks. These results suggest that connecting evidence is a transferable training target, and thus tasks whose training data are costly to build can benefit from data that are far easier to construct and verify.

AI USE STATEMENT

We used generative AI models within EVIDENCECONNECT for region discovery, question and distractor generation, answer verification, and difficulty filtering. These models include Qwen3.5-397B-A17B, Kimi-K2.6, GLM-5.2, and Qwen3.6-35B-A3B. Their roles and sampling settings are documented in Appendix A, and the prompt templates are provided in Appendix B.

In addition to these research uses, we used generative AI tools to assist with manuscript writing and language polishing, including improving wording, clarity, and readability. The authors retain responsibility for the final manuscript, including its methods, experimental results, and scientific claims.

ETHICS STATEMENT

The images used to construct our synthetic training data are drawn exclusively from SA-1B (Kirillov et al., 2023), which is publicly available for research under the SA-1B Dataset Research License. According to the dataset documentation, the source images were licensed by the dataset providers, and faces and license plates were de-identified before release. Use and redistribution of the underlying images remain subject to the applicable dataset license and permissions.

Our training annotations are generated automatically and may inherit errors and biases from both the source imagery and the generative models. Cross-model verification and region-ablation tests are intended to reduce annotation errors, but do not guarantee correctness or eliminate bias. Downstream applications should account for these limitations and assess potential privacy and fairness implications in their intended deployment settings.

REFERENCES

Amirhossein Abaskohi, Spandana Gella, Giuseppe Carenini, and Issam H Laradji. Fm2ds: Few-shot multimodal multihop data synthesis with knowledge distillation for question answering. *arXiv preprint arXiv:2412.07030*, 2024.

Hanbo Bi, Zhiqiang Yuan, Zexi Jia, Jiapei Zhang, Chongyang Li, Peixiang Luo, Ying Deng, Xiaoyue Duan, and Jinchao Zhang. F2RVLM: Boosting fine-grained fragment retrieval for multi-modal long-form dialogue with vision language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain 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*, pp. 14455–14465, 2024a.

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Yiping Bao, et al. Babyvision: Visual reasoning beyond language. *arXiv preprint arXiv:2601.06521*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

Anurag Das, Adrian Bulat, Alberto Baldrati, Ioannis Maniadis Metaxas, Bernt Schiele, Georgios Tzimiropoulos, and Brais Martinez. More images, more problems? a controlled analysis of vlm failure modes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7423–7442, 2026.

DeepSeek-AI. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Suhaas Garre, Chris Mutty, Sushant Mehta, and Edwin Chen. Chartography: A benchmark for professional chart understanding. *arXiv preprint arXiv:2608.10677*, 2026.

Gemini Robotics Team. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6864–6890, 2024.

Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. *arXiv preprint arXiv:2308.16463*, 2023.

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)*, pp. 6700–6709, 2019.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. MANTIS: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

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*, 123:32–73, 2017. doi: 10.1007/s11263-016-0981-7.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Youngwan Lee, Soojin Jang, Yoorhim Cho, Seunghwan Lee, Yong-Ju Lee, and Sung Ju Hwang. Multihopspatial: Multi-hop compositional spatial reasoning benchmark for vision-language model. In *European Conference on Computer Vision*, pp. 580–598. Springer, 2026.

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. *arXiv preprint arXiv:2505.21500*, 2025a.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive training for spatial reasoning in vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Kaican Li, Lewei Yao, Jiannan Wu, Tiezheng Yu, Jierun Chen, Haoli Bai, Lu Hou, Lanqing Hong, Wei Zhang, and Nevin L Zhang. Insight-o3: Empowering multimodal foundation models with generalized visual search. In *The Fourteenth International Conference on Learning Representations*, 2025b.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8778–8786, 2025c.

Xiping Li and Jianghong Ma. AIM-CoT: Active information-driven multimodal chain-of-thought for vision-language reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26656–26681, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1227.

Shijie Lian, Changti Wu, Laurence Tianruo Yang, Hang Yuan, Bin Yu, Lei Zhang, and Kai Chen. Euclid's gift: Enhancing spatial perception and reasoning in vision-language models via geometric surrogate tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9824–9835, 2026.

Zichao Lin, Yifeng Xie, Bowen Qu, Haiming Wang, Jia Li, Haoning Wu, Yuhao Dong, Zuhao Yang, Jinguo Zhu, Haoyu Lu, et al. Perceptionbench: Evaluating atomic visual perception in multimodal large language models. *arXiv preprint arXiv:2607.24957*, 2026.

Xianjie Liu, Yiman Hu, Yixiong Zou, Liang Wu, Jian Xu, and Bo Zheng. Hide: Rethinking the zoom-in method in high resolution mllms via hierarchical decoupling. *arXiv preprint arXiv:2510.00054*, 2025.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. MMDU: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for LVLMs. *arXiv preprint arXiv:2406.11833*, 2024.

Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Yongbin Li, Minzheng Wang, Pengpeng Zeng, Lianli Gao, Heng Tao Shen, Yunshui Li, Hamid Alinejad-Rokny, Xiaobo Xia, Jingkuan Song, and Fei Huang. MMEvol: Empowering multimodal large language models with evol-instruct. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19655–19682. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1009.

Shravan Nayak, Xiangru Jian, Kevin Qinghong Lin, Juan A. Rodriguez, Montek Kalsi, Rabiul Awal, Nicolas Chapados, M. Tamer Özsu, Aishwarya Agrawal, David Vazquez, Christopher Pal, Perouz Taslakian, Spandana Gella, and Sai Rajeswar. Ui-vision: A desktop-centric gui benchmark for visual perception and interaction. *arXiv preprint arXiv:2503.15661*, 2025.

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

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

Xiyu Ren, Zhaowei Wang, Yiming Du, Zhongwei Xie, Chi Liu, Xinlin Yang, Haoyue Feng, Wenjun Pan, Tianshi Zheng, Baixuan Xu, et al. Memlens: Benchmarking multimodal long-term memory in large vision-language models. *arXiv preprint arXiv:2605.14906*, 2026.

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. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen, Shilin Yan, Hongwei Xue, Shuaiqi Lu, Xiaojun Tang, Guannan Zhang, Tiancheng Zhao, and Jianwei Yin. Mm-condchain: A programmatically verified benchmark for visually grounded deep compositional reasoning. *arXiv preprint arXiv:2603.12266*, 2026.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024.

Zhaochen Su, Jincheng Gao, Hangyu Guo, Zhenhua Liu, Lueyang Zhang, Xinyu Geng, Shijue Huang, Peng Xia, Guanyu Jiang, Cheng Wang, et al. Agentvista: Evaluating multimodal agents in ultra-challenging realistic visual scenarios. *arXiv preprint arXiv:2602.23166*, 2026.

Junyoung Sung, Seungwoo Lyu, Minjun Kim, Sumin An, Arsha Nagrani, and Paul Hongsuck Seo. CRIT: Graph-based automatic data synthesis to enhance cross-modal multi-hop reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jayant Sravan Tamarapalli, Rynaa Grover, Nilay Pande, and Sahiti Yerramilli. CountQA: How well do MLLMs count in the wild? *arXiv preprint arXiv:2508.06585*, 2025.

Duong T Tran, Trung-Kien Tran, Manfred Hauswirth, and Danh Le Phuoc. Reasonvqa: A multi-hop reasoning benchmark with structural knowledge for visual question answering. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18793–18803. IEEE, 2025.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025a.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Mindcube: Spatial mental modeling from limited views, 2026. *arXiv preprint arXiv:2506.21458*, 2025b.

Shenzhi Wang, Shixuan Liu, Jing Zhou, Chang Gao, Xiong-Hui Chen, Binghai Wang, An Yang, Shiji Song, Bowen Yu, Gao Huang, et al. Hopchain: Multi-hop data synthesis for generalizable vision-language reasoning. *arXiv preprint arXiv:2603.17024*, 2026.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025c.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, Yangqiu Song, and Mark Steedman. MMLongBench: Benchmarking long-context vision-language models effectively and thoroughly. In *Advances in Neural Information Processing Systems*, 2025d.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. First sft, second rl, third upt: Continual improving multi-modal llm reasoning via unsupervised post-training. *Advances in Neural Information Processing Systems*, 38:62293–62318, 2025.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a.

Lai Wei, Chengqi Li, Jiapeng Li, Ruina Hu, Yue Wang, and Weiran Huang. Clbench-v: Evaluating multimodal context learning from grounding to knowledge acquisition. *arXiv preprint arXiv:2607.25294*, 2026b.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094, June 2024.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *International Conference on Learning Representations (ICLR)*, 2025.

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

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, 2025.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *Advances in Neural Information Processing Systems*, 38, 2026a.

Junkai Zhang, Yihe Deng, Kai-Wei Chang, and Wei Wang. Thinking with visual grounding. *arXiv preprint arXiv:2606.16122*, 2026b.

Letian Zhang, Quan Cui, Bingchen Zhao, and Cheng Yang. Oasis: One image is all you need for multimodal instruction data synthesis. *arXiv preprint arXiv:2503.08741*, 2025a.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Bing Li, Chunfeng Yuan, Guangting Wang, Fengyun Rao, Ying Shan, and Weiming Hu. MMhops-R1: Multimodal multi-hop reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026c.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025b.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations (ICLR)*, 2025c.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Henry Hengyuan Zhao, Pan Zhou, and Mike Zheng Shou. GENIXER: Empowering multimodal large language models as a powerful data generator. In *Computer Vision – ECCV 2024*, volume 15081 of *Lecture Notes in Computer Science*, pp. 129–147. Springer, 2024.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Appendix

A Implementation Details

This section complements Section 3 with the concrete settings used to construct our training data, following the order of that section. The pipeline mainly uses Qwen3.5-397B-A17B (Qwen Team, 2026a), and Table 4 summarizes all model calls, each with the thinking mode disabled.

Table 4: Model calls in the EvidenceConnect pipeline. V_G denotes the privileged layout, V_G^{-i} the layout with region r_i masked, and V_I the policy view.

Stage	Model	Visual input	Calls	Temp.
Evidence discovery	Qwen3.5-397B-A17B	source image	1 per image	0.4
Evidence grouping	Qwen3.5-397B-A17B	none (text only)	1 per image	0.3
Question realization	Qwen3.5-397B-A17B	V_G	1 per group	0.5
Answer anchoring	Qwen3.5-397B-A17B, Kimi-K2.6	V_G	up to 4 + 4 per candidate	0.4
Equivalence judging	GLM-5.2	none (text only)	when needed	0.0
Drop-one testing	Qwen3.5-397B-A17B	V_G^{-i}	1–3 per region	0.4
Unmarked rewrite	Qwen3.5-397B-A17B	source image	1 per aggregate question	0.2
Unmarked audit	Qwen3.5-397B-A17B, Kimi-K2.6	source image	2 + 2 per rewrite	0.0
Distractor generation	Qwen3.5-397B-A17B	marked image	1 per marked candidate	0.8
Difficulty filtering	Qwen3.6-35B-A3B	V_I	4 per example	0.7

Image source and pre-filtering. We use SA-1B (Kirillov et al., 2023) as the sole image source. Before any model call, we keep only images whose longer side has at least 1280 pixels, whose mean per-channel intensity entropy is at least 6.5 bits, whose Canny edge density is at least 0.04, and whose luminance standard deviation is at least 60.

Grounded evidence discovery. For each image, a single call with the image and its caption (Prompt B.1) judges whether the image can support multi-region questions and returns up to 18 region proposals, whose box coordinates are converted to pixels and clamped to the image. We then discard boxes that cover less than 0.04% or more than 18% of the image, as well as boxes whose IoU with an earlier kept box is at least 0.65. Images judged unsuitable, or with fewer than four remaining regions, are skipped.

Task-conditioned evidence grouping. A text-only call receives the caption together with the label and local description of each region, ranked by grounding confidence (Prompt B.2), and proposes up to six groups of 3–7 regions. Each group is annotated with one of 22 task families, which map to four necessity modes that refine the three question forms of Section 3.1. Specifically, `answer_sensitive` covers aggregation (counting, numeric aggregation, and arithmetic), `set_complete` (filtering, anomaly, extremum, matching, and geometric, topological, or depth relations) and `pair_complete` (cross-region pairing) cover comparison and matching, and `chain_dependency` (OCR, spatial, viewpoint, conditional, Boolean, relational, and multi-hop chains) covers dependency chains and compositions. We keep groups whose region identifiers are valid, whose size lies in the allowed range, and that do not repeat an earlier group. In addition, the candidates of a filtering, anomaly, or extremum group must not mix coarse semantic classes, and groups with distinct task families are kept before groups that repeat a family.

Privileged evidence layout. Each region is cropped at native resolution with a margin of 8% of its box size on every side. The crops are placed on a white canvas through an order-preserving remapping of each image axis, which removes the rows and columns not covered by any crop, and the canvas is uniformly downscaled only when its longer side exceeds 4096 pixels. Each region receives a distinct color from a seven-color palette, which bounds the group size at seven. The same colors are used in the question text, in the privileged and ablated layouts, and for the outlines on the marked policy view.

Question realization. For each group, the generator receives the privileged layout, the group metadata (labels, local descriptions, natural referents, box coordinates, and outline colors), four

blueprints sampled from a library of 104 reasoning templates, and up to four additional free-form slots, and it returns up to eight candidates (Prompt B.3). The first blueprint follows the proposed task family when possible, and the remaining ones diversify answer forms and question types. Each candidate declares its question type, necessity mode, answer semantics (scalar, unordered set, or ordered pair), answer kind (outline color, number, Boolean pattern, semantic phrase, text, semantic set, or semantic pair), the contribution of every region, the predicted answer after removing each region, and a reasoning certificate. Open-vocabulary answers must contain fewer than ten words.

Rule-based checks and candidate selection. A certificate is valid only if every step reads from earlier steps alone, every selected region lies on a path to the output step, and no step is disconnected from it. Aggregation, arithmetic, and Boolean questions must further use an executable certificate with a restricted operator set. For aggregation questions, every region must receive a positive value, and the values must sum to the answer without all being one, since the answer would otherwise equal the number of outlined regions. Similarly, arithmetic expressions must reference every region and evaluate to the answer. We further reject candidates that expose region identifiers, do not refer to the marked regions, use conditional wording without explicit branches or outside conditional types, count vague background targets, request more than one output for a scalar answer, omit an explicit coordinate frame for spatial types, leak the answer through an outline color, use subjective phrasing, or declare an answer form that the answer does not match. Candidates whose own predicted answer after removing some region equals their proposed answer are rejected as well. The remaining candidates are ranked by the generator's self-assessed naturalness, visual confidence, and multi-region scores, with bonuses for compositional types and non-numeric answers, and this ranking only determines which candidates proceed to verification. Per group, we keep up to four top-ranked candidates, preferring distinct question types and allowing at most one numeric and two outline-color answers. Per image, we keep up to eight candidates but always retain the best candidate of each question type.

Cross-model answer anchoring. Each candidate is answered on the privileged layout by both anchoring families, with up to four samples per family and without the region descriptions or the generator's answer (Prompt B.4). For closed-form questions, the prompt also lists the allowed answer labels, and we first draw two samples per family. Such a candidate is rejected early unless at least three of these four responses agree and each family contributes at least one of them. Responses are clustered by normalized string matching with numeric equivalence, and if the largest cluster has fewer than six members, the text-only judge re-clusters all responses by semantic equivalence (Prompt B.5). A candidate is accepted if the largest cluster is not an insufficiency response and contains at least three of the four responses from each family, and its answer then becomes $a^\star$. Finally, $a^\star$ is re-checked against the answer-form constraints above.

Adaptive drop-one testing. For each selected region, its crop in the privileged layout is replaced by a gray block with a red cross, while all other crops stay in place. The first anchoring family answers each ablated view once, with a mode-specific addendum to the answering prompt (Prompt B.6). If the response matches $a^\star$ or cannot be parsed, the same view is queried twice more, and the region is redundant if at least two of its three responses match $a^\star$. A candidate is rejected if two or more of its regions are redundant.

Policy view. Unmarked rewriting is attempted only for aggregate question types (object counts, attribute counts, and sums of visible numbers), whose answers do not depend on the outlines, whereas the answers of selection and pairing questions are often outline colors themselves. The first anchoring family rewrites the question for the unmarked image and proposes distractors (Prompt B.7), and rewrites that still mention boxes or outlines are discarded. Two audits from each family (Prompt B.8) must then all report unique references, task preservation, identifiable targets, multi-region dependence, and necessity of every evidence area. Each audit must also recover $a^\star$ and ground exactly one box per intended region, and the grounded boxes must match the intended regions one-to-one, where a grounded box matches a region if its center lies inside the region's box enlarged by 15%. Otherwise, the example keeps the marked policy view, on which the box of each selected region is outlined in its assigned color.

Multiple-choice conversion. For marked examples, the first anchoring family proposes three distractors given the marked image, the question, and $a^\star$ (Prompt B.9), whereas unmarked examples reuse the distractors from the rewrite step. We remove empty distractors, distractors that match $a^\star$ after normalization, duplicates, and, for closed-form answers, options outside the label vocabulary. If fewer than three distractors remain, we fill the missing options deterministically, for example with alternative subsets of set-valued answers, unused vocabulary labels, or neighboring integers for count answers. The four options are then shuffled, and the prompt asks for the option letter.

Difficulty filtering. The screening model answers each example four times on its policy view (Prompt B.10). Examples with four correct responses or any unparseable response are discarded, and the remaining ones are labeled easy, medium, hard, or very hard when three, two, one, or none of the responses are correct. Finally, we keep at most four examples per source image, and each training example consists of the policy view and the multiple-choice question.

B Prompt Templates for EvidenceConnect

This section lists the prompt templates used by EvidenceConnect. We replace instance-specific values with angle-bracketed placeholders and condense long templates for readability. Specifically, we keep their task definitions, key constraints, and main output fields, and omit examples, formatting reminders, repeated instructions, and fields unused in our setting.

B.1 Evidence Discovery and Question Construction

Prompt B.1: Grounded region inventory

```
[IMAGE]
You are a precise visual grounding annotator and data-quality gate.

Task: Decide whether this image can support natural questions whose
answer genuinely depends on 3--7 distinct regions, then propose up to
<MAX_REGIONS> useful evidence regions.

Image caption, if available: <IMAGE_CAPTION>

Return regions that are:
- visually concrete and clearly localizable;
- useful for multi-region reasoning, including people, animals, objects,
signs, screens, text regions, fine details, and distinctive parts;
- preferably small or medium-sized, spatially diverse, and not overly
overlapping;
- suitable for coherent tasks such as positive-contribution counting,
closed-set filtering, anomaly/extremum detection, candidate matching,
pairing, OCR--visual verification, or a short dependent spatial chain.

Prioritize groups of related regions over isolated interesting objects.
The image is suitable only if at least one coherent all-region task is
possible. Avoid vague regions such as "background," "scene," "sky,"
"room," or the whole image.

For each region, return:
- label: short name;
- bbox: [x1,y1,x2,y2] in pixels, or normalized 0--1000 coordinates;
- caption: one concise visible description;
- natural_referent: an identifying phrase without coordinates or IDs;
- role: candidate | reference | comparator | text_region |
      count_target | fine_part | distractor;
- confidence: grounding confidence.
```

```
Return JSON only:
{
"suitable_for_multi_region_vqa": true,
"richness_reason": "...",
"coherent_sets": [
  {"labels": [...], "task_family": "..."}
],
"regions": [
  {
    "label": "...", "bbox": [...], "caption": "...",
    "natural_referent": "...", "role": "...", "confidence": 0.0
  }
]
}
```

Prompt B.2: Task-conditioned region-group proposal

```
You are an expert VQA data curator. Below are <N> candidate evidence
regions detected in one image.

Image caption: <IMAGE_CAPTION>
Regions:
- <REGION_ID>: label=<LABEL>; caption=<LOCAL_CAPTION>
- ...

Propose up to <MAX_GROUPS> combinations that are good for natural
questions requiring many regions. A good combination is a coherent
candidate/evidence set, not a bag of unrelated boxes. Prefer the
largest valid set, but every selected region must have a real role.

Allowed task families and necessity modes:
- answer_sensitive: aggregate_count, aggregate_attribute_count,
aggregate_numeric, arithmetic_composition;
- set_complete: set_filtering, anomaly_detection, extremum,
candidate_matching, geometric_line_relation, topological_relation,
depth_ordering;
- pair_complete: cross_region_pairing;
- chain_dependency: ocr_visual_verification, spatial_chain,
dependent_chain, viewpoint_transform, agent_relative_direction,
spatial_transitivity, conditional_composition, boolean_composition,
relational_composition, multi_hop_reasoning.

Selection rules:
1. Select 3--7 regions and prefer larger groups only when coherent.
2. Every region must change an aggregate, belong to the closed
 candidate/pair set, or supply an indispensable chain constraint.
3. Prefer objective text, numbers, countable objects, repeated classes,
 comparable parts, and unambiguous attributes.
4. Filtering, anomaly, and extremum candidates must be comparable.
5. Every aggregate-count region must contribute a positive count.
6. Spatial spread is preferred, but coherence has priority.
7. Diversify task families and do not default to counting.
8. Skip the image if no valid combination exists.

Return JSON only:
[
{
  "region_ids": ["r000", "r001", "r003"],
  "task_family": "anomaly_detection",
```

```
  "necessity_mode": "set_complete",
  "reason": "short objective rationale"
}
]
```

Prompt B.3: Privileged question and certificate realization

```
[PRIVILEGED LAYOUT WITH K COLOR-OUTLINED CROPS]
Create <NUM_QUESTIONS> diverse multi-region VQA questions.

Scope contract:
- The selected set is exactly <REGION_IDS>.
- Every question must genuinely use all K regions.
- evidence_region_ids, region_contributions, answer_if_dropped, and
the reasoning certificate must cover exactly this set.
- Never add decorative observations for otherwise unused regions.
- Questions must be answerable from visible pixels in this layout,
not from region captions.

Visible crop anchors (the same outline colors appear on the marked image):
- <REGION_ID>: label=<LABEL>; outline_color=<COLOR>; bbox=<BBOX>;
  caption=<LOCAL_CAPTION>; natural_referent=<REFERENT>
- ...

Task guidance:
- Preferred task family: <TASK_FAMILY>.
- Preferred necessity mode: <NECESSITY_MODE>.
- Selected reasoning blueprints: <TEMPLATE_SPECS>.
- Instantiate a blueprint only when its facts are objectively visible.
- Include diverse closed-set, comparison/matching, OCR/spatial/
relational, and compositional structures.
- Prefer select-then-report questions in which several regions identify
a target and a second visible property supplies the final answer.
- Keep the answer short, objective, unique, and absent from the stem.

Necessity contract:
- answer_sensitive: every region contributes a positive value, so
masking it changes the result or makes the aggregate unknown.
- set_complete / pair_complete: all regions form the explicitly closed
candidate or pairing set; masking one prevents proving completeness.
- chain_dependency: every region supplies an indispensable clue or hop.
- For every region, answer_if_dropped must differ from the reference
answer or be "insufficient information."

Certificate contract:
- First design a natural question from pixels; then provide its
certificate. Do not verbalize the certificate mechanically.
- Use mode="executable" for deterministic reading, counting, filtering,
matching, comparison, arithmetic, Boolean, and lookup programs.
- Use mode="evidence_graph" for open-vocabulary semantic or unusual
multi-hop reasoning.
- Every step has id, op, regions, inputs, and query.
- Inputs may reference only earlier steps.
- Every selected region must lie on a path to output_step.
- No disconnected or decorative step is allowed.

Wording contract:
- Ask for one final answer.
- Use one explicit coordinate frame for spatial questions.
```

```
- Use outline colors only as reference anchors, never as leaked object
attributes.
- Open answers must be objective and fewer than 10 words.
- Do not use subjective comparisons, vague counts, or invisible facts.

Return JSON only, one object per candidate:
[
{
  "template_id": "...",
  "question": "...",
  "question_type": "...",
  "necessity_mode": "answer_sensitive | set_complete |
                     pair_complete | chain_dependency",
  "answer_semantics": "scalar | unordered_set | ordered_pair",
  "final_answer_kind": "anchor | numeric | boolean | semantic |
                        semantic_set | semantic_pair | text",
  "evidence_region_ids": [...],
  "region_contributions": {"r000": "...", "...": "..."},
  "region_values": {"r000": 0},
  "aggregation": "...",
  "operation_expression": "...",
  "candidate_region_ids": [...],
  "closed_candidate_set": true,
  "pair_region_ids": [[...], ...],
  "chain_region_ids": [...],
  "coordinate_frame": "image_plane | camera | agent | object |
                       scene_3d | not_applicable",
  "answer_vocabulary": [...],
  "answer_if_dropped": {"r000": "...", "...": "..."},
  "hypothetical_answer": "...",
  "reasoning_certificate": {
    "mode": "executable | evidence_graph",
    "steps": [
      {"id": "s1", "op": "...", "regions": ["r000"],
       "inputs": [], "query": "..."}
    ],
    "output_step": "..."
  },
  "naturalness_score": 1,
  "visual_confidence": 1,
  "multi_region_score": 1
}
]
```

B.2 Answer Anchoring and Dependency Checks

Prompt B.4: Privileged-view answer anchoring

```
[PRIVILEGED LAYOUT]
Answer the question using only the provided image.

Question: <QUESTION>
Allowed canonical answer labels: <OPTIONAL_NEUTRAL_VOCABULARY>

You may reason briefly before answering. Put only the final short answer
inside <answer>...</answer>.

Rules:
- Use an objectively supported expression of fewer than 10 words.
```

```
- Use Arabic digits for numbers, with no units or separators.
- Do not emit coordinates, grounding tokens, or box markers.
- Use the same language as the question.
- If evidence is insufficient, output
<answer>insufficient information</answer>.
- The answer tag is required; responses without it are discarded.
```

Prompt B.5: Semantic-equivalence clustering

```
You are an answer-equivalence judge for a VQA task.

Question: <QUESTION>
Answers:
A1: <ANSWER_1>
...
AN: <ANSWER_N>

Group answers by semantic equivalence. Equivalent paraphrases belong
together; answers with different values, attributes, or truth values
must remain separate.

Return strict JSON only:
{
"groups": [["<CANONICAL_FORM>", [<ANSWER_INDICES>]], ...],
"largest_group_size": <INTEGER>,
"canonical": "<CANONICAL_FORM_OF_LARGEST_GROUP>"
}

Every answer must occur in exactly one group. Break size ties using the
lexicographically smallest canonical form.
```

Prompt B.6: Mode-specific drop-one verification The drop-one verifier inserts one of the following addenda into Prompt B.4, depending on the necessity mode.

```
[answer_sensitive]
ABLATION VIEW: one required aggregate region has been replaced by a
gray masked box. If the full aggregate cannot be computed, answer
"insufficient information"; do not count only visible boxes.

[set_complete or pair_complete]
ABLATION VIEW: one colored candidate box has been replaced by a gray
masked box. The masked box remains a member of the explicitly closed
candidate/pairing set. If its hidden content prevents proving the
complete subset, unique anomaly, global extremum, unique match, or
complete pairing, answer "insufficient information."

[chain_dependency]
ABLATION VIEW: one required clue region has been replaced by a gray
masked box. If the missing clue breaks the reasoning chain, answer
"insufficient information"; do not guess.
```

B.3 Policy-View and Multiple-Choice Construction

Prompt B.7: Unmarked-question rewrite

```
[UNMARKED ORIGINAL IMAGE]
You are given an anchored visual question and its verified answer.

Anchored question: <QUESTION>
```

```
Verified answer: <ANSWER>

Create a minimally changed question that works on the unmarked image.
Replace box/outline references with visually unique natural descriptions
or a clearly exhaustive image-level scope. Do not add facts, change the
operation, or change the answer. If unique reference cannot be
guaranteed, set safe_without_marks=false. Also generate three plausible
incorrect distractors in the same answer format.

Return JSON only:
{
"safe_without_marks": true,
"natural_question": "...",
"distractors": ["...", "...", "..."]
}
```

Prompt B.8: Independent audit of the unmarked rewrite

```
[UNMARKED ORIGINAL IMAGE]
Audit this visual question using only the unmarked original image.

Anchored source question: <ANCHORED_QUESTION>
Unmarked natural question: <NATURAL_QUESTION>

Independently answer the question and judge whether every reference or
exhaustive target set is uniquely determined. The question must require
multiple spatially separated regions. Ground each evidence area with a
box in normalized 0--1000 coordinates, locating it from pixels rather
than inferring it from wording. Also judge whether the rewrite preserves
the source task and whether every grounded area is required.

Return JSON only:
{
"answer": "...",
"reference_status": "unique | ambiguous",
"same_task_as_source": true,
"all_targets_identifiable": true,
"requires_multiple_regions": true,
"every_evidence_area_required": true,
"evidence_boxes": [
  {"description": "...", "bbox_2d": [x1,y1,x2,y2]}
],
"reason": "..."
}
```

Prompt B.9: Multiple-choice distractor generation

```
[MARKED ORIGINAL IMAGE]
You are given a visual question and its correct answer.

Original question: <QUESTION>
Correct answer: <ANSWER>

Generate three plausible but incorrect multiple-choice distractors.
Do not rewrite, paraphrase, simplify, or neutralize the question.
Keep the stem exactly unchanged. Distractors must be visually relevant,
distinct from the answer, and matched in answer format and approximate
length. For open-vocabulary answers, use concrete alternatives with the
same grammatical form and specificity. Do not use "none of the above,"
```

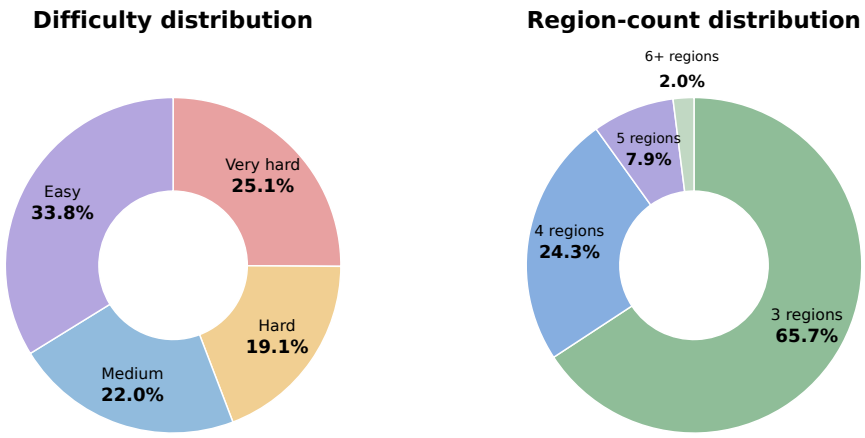


Figure 5: Composition of the 20.4K training examples by the number of selected regions per example and by difficulty level.

```
"all of the above," or "cannot be determined."

Return JSON only:
[{"distractors": ["d1", "d2", "d3"]}]
```

Prompt B.10: Difficulty sampling on the policy view

```
[POLICY VIEW: MARKED OR UNMARKED IMAGE]
Answer this multiple-choice visual question using only the image.

<QUESTION_AND_OPTIONS>

You may reason briefly before answering. The final answer must be:
<answer>X</answer>

Rules:
- X must contain only A, B, C, or D.
- Do not output option text, coordinates, grounding tokens, or markers.
- Keep any reasoning to a few short sentences.
- The answer tag is required; otherwise the response is discarded.
```

C Training Data Statistics

Our training set contains 20.4K examples synthesized from SA-1B images by the pipeline in Section 3 and Appendix A, and Figure 5 summarizes its composition. Each example involves three to seven regions, most often three (65.74%) or four (24.34%), followed by five (7.93%) and six or seven (1.99%). The difficulty levels, defined by how many of the four screening responses are correct, are broadly distributed. Specifically, easy, medium, hard, and very hard examples account for 33.75%, 22.04%, 19.07%, and 25.14% of the data, so more than 44% of the examples are hard or very hard. Very hard examples, which the screening model never answers correctly, are kept, since their reference answers have already been anchored by two stronger VLMs.

D TRAINING DETAILS

All models are trained with the same GRPO (Shao et al., 2024) configuration for two epochs, with a batch size of 128 and a learning rate of 10^{-6}. For each prompt, we sample $G{=}8$ rollouts at temperature 0.7 with a maximum response length of 32K tokens, and the thinking mode is disabled during rollouts to keep responses concise. We use decoupled clipping ratios $\epsilon_{\text{low}}{=}0.2$ and $\epsilon_{\text{high}}{=}0.3$, as proposed in DAPO (Yu et al., 2025), and remove the KL penalty ($\beta_{\text{KL}}{=}0$), so that the policy is not anchored to the base model. Finally, the reward checks whether the predicted option letter matches the ground-truth option.

E ATTENTION MAP COVERAGE ANALYSIS

Evaluation setting. We compare PerceptVLM-4B/9B with their Qwen3.5 base models on the same 405 TreeBench examples. Within each scale, both checkpoints use the base-model processor and identical image–question pairs, visual-token grids, and evidence annotations. Since Qwen3.5 interleaves linear-attention and full-attention layers, we evaluate all eight full-attention layers, $\mathcal{L} = \{4, 8, 12, 16, 20, 24, 28, 32\}$ (1-based).

Attention extraction. For each example, we apply the chat template with the generation prompt and perform a single forward pass in evaluation mode using eager attention, without generating an answer. We extract attention from the final token of this complete input sequence to the image tokens. This position predicts the first answer token, so its attention indicates where the model looks as it begins to answer.

Let V_i denote the visual-token indices for example i, and let q_i be its final input-token position. For model m at layer ℓ, let $T^m_{i,\ell,h}(q_i, v)$ denote the attention weight returned by head h from q_i to visual token v. We convert the extracted weights to float32, average equally over all $n^{m,\ell}_{\text{heads}}$ heads, and then renormalize within the image tokens:

$$
\begin{aligned}
u^m_{i,\ell}(v) &= \frac{1}{n^{m,\ell}_{\text{heads}}} \sum_{h=1}^{n^{m,\ell}_{\text{heads}}} T^m_{i,\ell,h}(q_i, v), \\
a^m_{i,\ell}(v) &= \frac{u^m_{i,\ell}(v)}{\sum_{v' \in V_i} u^m_{i,\ell}(v')}.
\end{aligned}
\tag{1}
$$

This renormalization over image tokens is applied once after head averaging, without renormalizing each head separately or applying an additional softmax.

Evidence mask. The visual tokens form an $H_i \times W_i$ grid for an image of size $H^{\text{img}}_i \times W^{\text{img}}_i$. An annotated box (x_0, y_0, x_1, y_1) covers columns $[\lfloor x_0 W_i / W^{\text{img}}_i \rfloor, \lceil x_1 W_i / W^{\text{img}}_i \rceil)$ and the analogous row range, clipped to the grid boundaries. This range includes every visual-token cell that overlaps the box. The evidence set $E_i \subseteq V_i$ is the union of the tokens covered by all annotated boxes, counting overlapping tokens once. All evaluated examples have $|E_i| > 0$.

Top-k evidence IoU. For each example, we set $k_i = |E_i|$ and select the k_i visual tokens with the largest head-averaged weights $u^m_{i,\ell}(v)$ to form $A^m_{i,\ell}$. Ranking by the normalized weights $a^m_{i,\ell}(v)$ gives the same ordering because the normalization divides all weights by a common positive value. Ties are broken by visual-token index in row-major order, so exactly k_i tokens are selected even when ties occur at the boundary. Because k_i is fixed by the annotation, the base and trained models receive the same token budget.

With $h^m_{i,\ell} = |A^m_{i,\ell} \cap E_i|$, we compute

$$
\text{IoU}^m_{i,\ell} = \frac{|A^m_{i,\ell} \cap E_i|}{|A^m_{i,\ell} \cup E_i|} = \frac{h^m_{i,\ell}}{2k_i - h^m_{i,\ell}}.
\tag{2}
$$

The score lies in $[0, 1]$, with larger values indicating better spatial alignment with the annotated evidence. Since $|A^m_{i,\ell}| = |E_i|$, precision equals recall, $p = h^m_{i,\ell}/k_i$, and the IoU equals $p/(2-p)$.

Table 5: **Top-k evidence IoU on TreeBench.** Mean IoU is reported as a percentage; Δ is PerceptVLM minus Qwen3.5 in percentage points. Each entry averages the same 405 paired examples. Differences are computed before rounding.

	4B			9B		
Layer	Qwen3.5	PerceptVLM	Δ	Qwen3.5	PerceptVLM	Δ
4	11.625	11.877	+0.252	10.622	10.699	+0.077
8	10.355	10.402	+0.047	10.069	10.051	-0.018
12	11.568	11.692	+0.124	11.073	11.050	-0.023
16	10.400	10.650	+0.251	11.080	11.256	+0.175
20	11.132	11.386	+0.255	11.493	11.777	+0.284
24	10.924	11.162	+0.238	10.890	11.242	+0.352
28	11.181	11.152	-0.029	11.147	11.680	+0.533
32	11.244	11.424	+0.179	10.471	10.860	+0.390

Aggregation and layer-wise results. We take the unweighted mean of the per-example IoU over the $N = 405$ paired examples, rather than pooling intersections and unions, and report differences in percentage points:

$$\overline{\mathrm{IoU}}_\ell^m = \frac{1}{N}\sum_{i=1}^{N} \mathrm{IoU}_{i,\ell}^m, \qquad \Delta_\ell = 100\left(\overline{\mathrm{IoU}}_\ell^{\mathrm{PerceptVLM}} - \overline{\mathrm{IoU}}_\ell^{\mathrm{Qwen3.5}}\right). \tag{3}$$

Table 5 shows gains in 7 of 8 layers at 4B and 6 of 8 layers at 9B. The exceptions, layer 28 at 4B and layers 8 and 12 at 9B, each decrease by less than 0.03 percentage points. These comparisons describe changes in sample means. The metric measures spatial alignment at the final prompt position, rather than total attention mass or answer correctness.

F Case Study

To complement the quantitative evaluation, we examine three selected examples in which PerceptVLM-4B answers correctly while its base model, Qwen3.5-4B, fails. Figure 6 presents examples from MIMIC, ZoomBench, and CountQA, covering cross-image category identification, spatially specified text recognition, and object counting. We analyze the models' visible responses to identify how their answers differ in the use of visual evidence and the satisfaction of question constraints.

Cross-image consistency. In Figure 6(a), the MIMIC question asks for a category that appears in all five images. The base model selects *wine glass*, even though its response explicitly acknowledges that this object is absent from some images. It justifies the answer by referring to the most prominent shared item across the majority of the set, effectively replacing the requirement of universal presence with majority overlap. PerceptVLM checks candidate categories against the full set of images. Its response rules out *wine glass* because it is absent from Images 4 and 5, and identifies *person* as the category shared by all five images. This example illustrates the importance of preserving the question's scope when integrating evidence across images: an object shared by several images does not necessarily satisfy a constraint over the entire set.

Question-conditioned text localization. Figure 6(b) asks for the black text above the lower archway. The base model recognizes the word *SEGATORI* on the adjacent building, but associates the question with inscriptions on the fountain and ultimately selects *FONTE*. Its response therefore reveals a mismatch between the recognized text and the region specified by the question. PerceptVLM identifies the entrance of the yellow building as the relevant lower archway and reads the sign immediately above it, yielding the correct answer, *SEGATORI*. The case highlights the role of spatial reference resolution in visual text understanding: recognizing a visible word must be accompanied by establishing whether it belongs to the queried location.

(a) MIMIC: Shared category



Image 1 Image 2 Image 3 Image 4 Image 5

Identify the class that appears in all these images.

Base: [...] While not present in every single image, it is the most prominent shared item across the majority of the set.
Answer: wine glass ✗

Ours: [...] Wine glass: No (4 and 5 don't have it). [...] So the only consistent class across all is person.
Answer: person ✓

Error analysis: Cross-image constraint violation. The answer must occur in every image.

(b) ZoomBench: Target text



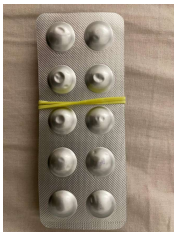
What is the text written in black above the lower archway?
A. FONTE B. SEGATORI C. TREVI D. ROMA

Base: [...] there is a building with a sign that reads "SEGATORI" [...] not part of the fountain's inscription.
Answer: A. FONTE ✗

Ours: [...] On that building, just above the entrance (which is a lower archway), there is a sign with black text. That text reads: "SEGATORI".
Answer: B. SEGATORI ✓

Error analysis: Target binding error. Base reads SEGATORI but assigns the query to the fountain.

(c) CountQA: Tablet counting



How many tablets are there?

Base: [...] The top section (above the yellow rubber band) has 2 tablets. The bottom section (below the yellow rubber band) has 4 tablets. So, total tablets = 2 + 4 = 6
Answer: 6 ✗

Ours: [...] The blister pack is arranged in 2 columns and 5 rows. [...] Total = 5 + 5 = 10 tablets
Answer: 10 ✓

Error analysis: Visual undercounting. The pack has two columns of five tablets.

Base: Qwen3.5-4B. Ours: PerceptVLM-4B.
Excerpts from visible responses. [...] marks omissions. Answer fields are reformatted.

Figure 6: **Qualitative comparisons between Qwen3.5-4B and PerceptVLM-4B.** Three selected examples show correct predictions by PerceptVLM-4B where the base model fails: (a) identifying a category present in every image; (b) associating recognized text with the queried location; and (c) counting all instances in a structured arrangement. Responses are excerpted from the original outputs, with omissions marked by `[...]` and answer fields reformatted. The enlarged image region in (b) is provided for visualization only.

Structured visual counting. In Figure 6(c), the CountQA question asks for the number of tablets in a blister pack. The base model divides the image into regions above and below the yellow rubber band, reports two and four tablets, respectively, and returns $2 + 4 = 6$. Its arithmetic is consistent with these counts, so the error lies in enumerating the tablets. PerceptVLM describes the pack as two columns and five rows, counts five tablets in each column, and obtains the correct total of $5+5 = 10$. Here, the response uses the regular spatial arrangement to account for the complete set of tablets.

Together, these cases reflect the steps of connecting evidence: integrating facts over the full set of images in (a), binding recognized text to the queried location in (b), and accounting for every instance before aggregating in (c).

G LIMITATIONS AND FUTURE WORK

Our work has two main limitations. First, although our single-image supervision transfers to multi-image and agent tasks, it does not directly model temporal or cross-frame dependencies, so extending the synthesis to evidence spread across video frames is a natural next step. Second, we apply RLVR to models of up to 27B parameters, and whether the same supervision benefits larger models or earlier training stages, such as pre-training, remains to be studied.