SELF-EVOLVING VISUAL QUESTIONER

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) are typically trained as passive answerers, while their ability to actively ask diverse, non-trivial, and visually grounded questions remains underexplored. Existing visual questioners' performance is bottlenecked by the availability of high-quality training data and the cost of curating it. We show that a VLM can iteratively improve itself as a visual questioner without external teachers or new human annotations. We propose a self-evolving framework that uses a VLM itself as a proposer, rewriter, and filter to produce harder, more informative, and visually grounded questions, while maintaining diversity across visual intents. These questions are then used to train the VLM in both questioner and answerer modes, with the updated questioner proposing questions for the next round. To evaluate the questioner, we introduce an evaluation protocol that assesses questions along perception, reasoning, and diversity dimensions. Experiments across three backbone VLMs show that our method substantially enhances question quality, increasing the mean question-generation score by approximately 78% after two rounds. Under the same training-data budget, our self-supervision is more effective than training on static source annotations. Moreover, the self-evolving questioner preserves its answering ability.[1]

1 INTRODUCTION

Vision-language models (VLMs) have achieved substantial progress on visual question answering, multimodal instruction following, and visual reasoning benchmarks. However, most are trained and evaluated primarily as passive answerers: questions are treated as fixed inputs provided by humans, curated datasets, or stronger external models, and learning focuses on producing better answers (Liu et al., 2023; Wu & Xie, 2024; Zhang et al., 2026). The complementary ability to ask informative and visually grounded questions remains comparatively underexplored. We view visual questioning as a fundamental capability for intelligent systems: it reflects how a model inspects an image, selects informative evidence, and formulates queries requiring meaningful perception and reasoning. The questions a model generates provide a concrete lens into this capability, as well as a source of supervision for improving it. Better visual questions can further expose richer evidence and reasoning paths, making them valuable supervision for improving visual questioning itself.

Despite its importance, existing visual question generation (VQG) methods remain fundamentally bottlenecked by static data distributions (Mi et al., 2024; Shen et al., 2024). Many approaches rely on human annotations, curated datasets, or stronger external models to provide supervision and consequently learn primarily to imitate existing question patterns (Zhao et al., 2024; Xie et al., 2025b; Zhang et al., 2025a). Such supervision shapes the visual content and question patterns a model learns to explore. Questions centered on salient objects or familiar templates may require little inspection of image-specific evidence, leaving richer spatial and contextual relationships unexplored. Improving the questioner therefore calls for supervision that broadens visual coverage and encourages more demanding, grounded reasoning. Continuously collecting this supervision from humans or stronger models, however, is costly and difficult to scale.

To move beyond this bottleneck, a natural alternative is to allow the model itself to construct better supervision through self-generated questions. However, simply scaling self-generated supervision does not automatically lead to self-improvement. Raw proposals can reinforce familiar question patterns and concentrate on a narrow range of visual evidence. The key challenge is to construct super-

[1] Code is available at this anonymous repository.

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.

vision that encourages broader exploration and more demanding reasoning while keeping questions grounded in image evidence.

To address this challenge, we propose a **self-evolving visual questioner** that improves QG through self-generated supervision while preserving answering ability. The central idea is to improve the questions used for supervision as the questioner learns. Given an image, the current model proposes candidate questions under multiple visual intents to explore different aspects of its content. Rewriting develops the proposals toward richer visual inspection and reasoning; filtering selects question–answer pairs that remain grounded and improve perception or reasoning difficulty. The selected pairs train the model in both QG and QA formats: QG supervision teaches it to generate questions and answers from an image, while QA supervision anchors its ability to answer supplied questions. The adapted model becomes the next-round proposer, closing the self-improvement loop over the same unlabeled image pool. The model constructs its own training supervision without external teachers, new human annotations, or auxiliary reward models.

To evaluate visual questioning beyond conventional QA accuracy, we introduce an evaluation protocol covering perception, reasoning, and diversity dimensions. Experiments across multiple VLM backbones show that our self-evolving framework substantially improves the visual-centric quality of generated questions while largely preserving QA performance. The reported across-backbone mean QG average increases by approximately 78%, with gains in visual search, evidence coverage, and contextual and spatial reasoning. These QG gains are accompanied by largely preserved answering ability. Ablations show how rewriting, filtering, and dual-format supervision contribute to these gains. Under a matched training-data budget, the constructed supervision also provides a stronger adaptation signal than directly sampled source annotations. Using the improved questions for downstream QA training further increases average accuracy under a fixed answer-generation model. Together, these results support self-evolution as a practical way to train stronger visual questioners from the model's own data.

Our contributions are summarized as follows:

- We propose an autonomous self-evolving framework for visual question generation that improves questioning through self-generated supervision while preserving answering ability.
- We combine multi-round proposal, rewriting, and filtering with dual-format training: refined questions teach the model to ask better questions, and the updated model constructs proposals for the next round.
- We introduce a structured evaluation protocol for perception, reasoning, and question-set diversity, enabling fine-grained analysis beyond aggregate QA accuracy.
- Experiments across three backbones establish substantial QG gains; component ablations and matched-budget comparisons demonstrate the value of refined self-supervision, including its use in downstream QA training.

2 From Answerer to Questioner

2.1 Problem Formulation

Given an image collection $\mathcal{I}$ and an initial VLM M_0, we seek to improve visual question generation (QG) from self-generated data while preserving question answering (QA). For image x, question q, and answer a, VQA predicts answers while QG predicts questions:

$$p_\theta(a \mid x, q) \quad \text{(VQA)}, \qquad p_\theta(q \mid x, c) \quad \text{(QG)}, \tag{1}$$

where c optionally specifies a question type, answer style, or reasoning intent to broaden coverage. Without this control, questions are generated directly from the image.

2.2 Visual Questioning Capability Evaluation Protocol

A central challenge in evaluating visual questions is that surface-level quality may not reflect visual informativeness. A question may be well-formed but shallow, ambiguous, weakly grounded in the image, or redundant with other questions for the same image. We therefore design an evaluation protocol to assess whether generated questions are visual-centric, challenging, and diverse.

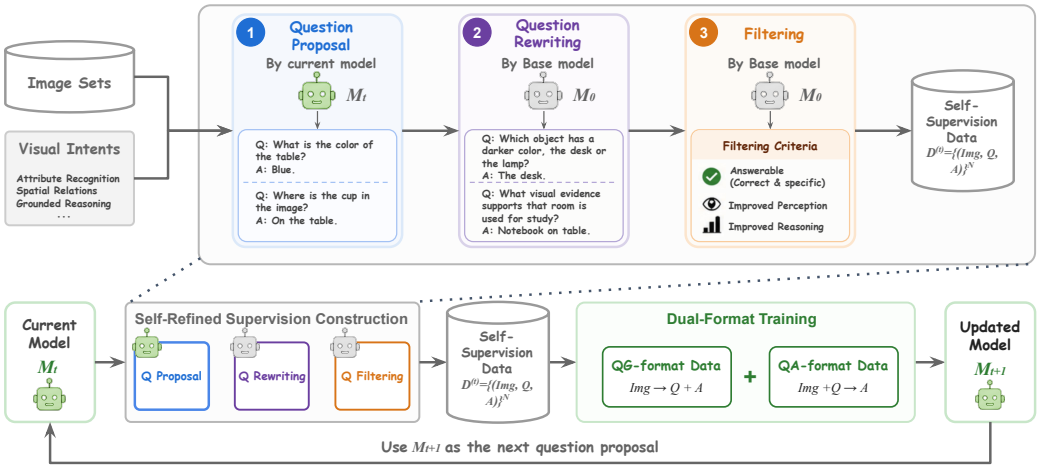

Figure 1: **Overview of our self-evolving visual questioner framework.** Given unlabeled image sets and visual-intent prompts, the current model M_t first proposes candidate questions. A stable base model M_0 then rewrites and filters these proposals to construct self-supervision data $\mathcal{D}^{(t)}$ with improved answerability, perception difficulty, and reasoning difficulty. The refined supervision is used for dual-format training with both QG-format and QA-format objectives, producing an updated model M_{t+1}. The updated model becomes the next-round proposer.

Our protocol evaluates generated questions at two levels. At the individual-question level, we measure **perception difficulty** and **reasoning difficulty**. Perception difficulty concerns what visual evidence must be identified to answer the question, and consists of two dimensions: *Visual Search Difficulty*, which measures how difficult it is to locate the required evidence, and *Visual Evidence Coverage*, which measures how broadly the question depends on evidence across the image. Reasoning difficulty concerns how the identified evidence must be interpreted, and consists of two dimensions: *Visual Context Reasoning*, which measures contextual interpretation from visible cues, and *Visual Spatial Reasoning*, which measures reasoning over spatial relations among image elements.

At the question-set level, we measure diversity through *Questioning Diversity*, which captures redundancy among questions generated for the same image. This set-level dimension is important because a question can be valid and reasoning-oriented in isolation, yet provide little additional supervision if it is a near-duplicate of other questions for the same image.

For individual-question evaluation, we use a GPT-based judge with clearly defined rubrics for the four perception and reasoning dimensions. The judge scores each question conditioned on the image, allowing the evaluation to focus on visual evidence and reasoning requirements rather than surface-level wording. For question-set diversity, we use a sentence embedding model to compute semantic distances among questions generated for the same image, measuring whether the question set provides non-redundant visual supervision. The GPT judge evaluates questions independently after training; its scores, feedback, and outputs are excluded from data construction and optimization. Detailed rubrics and diversity computation are provided in Appendix B.

3 Self-Evolving Visual Questioner

Our framework consists of two stages: self-supervision data construction and dual-format model training, aiming to improve questioning capability while preserving answering ability. Given unlabeled images, the VLM proposes candidate questions, rewrites them into harder, visually grounded questions, and filters the resulting pairs for validity and improvement. The retained question–answer pairs train the model in both questioner and answerer modes, and the adapted model becomes the proposer for the next round. This self-generated and self-refined supervision requires no external teachers or new human annotations. **Rewriting and filtering are used only during self-evolution training.** During evaluation, we directly assess the model's unmodified questions without additional rewriting or filtering, measuring its intrinsic questioning capability.

3.1 Question Proposal

For each image $x_i \in \mathcal{I}$, model M_t proposes a set of K questions, $\mathcal{Q}_i^{\text{prop}} = \{q_{i,1}^{\text{prop}}, \ldots, q_{i,K}^{\text{prop}}\}$. Here M_t is the base model or a checkpoint adapted in a previous round. To explore different aspects of each image, we prompt M_t with multiple visual intents: recognition, comparison, spatial relations, scene understanding, and grounded reasoning. These intents diversify the starting points for refinement; proposals must be answerable from visible image evidence. The same model generates each answer from the image and proposed question, yielding candidate triples:

$$\mathcal{C}_t^{\text{prop}} = \{(x_i, q_{i,j}^{\text{prop}}, a_{i,j}^{\text{prop}})\}_{i,j}. \tag{2}$$

3.2 Question Rewriting

A key design choice is to separate question proposal from question rewriting. The evolving proposer M_t determines which questions to explore as its questioning behavior changes through training. Rewriting starts from these concrete proposals and asks for finer visual inspection, broader evidence, or an additional reasoning step. This gives refinement a specific visual target to improve.

Using the adapted checkpoint to rewrite its own proposals couples proposal and refinement to the same weight updates. We therefore instantiate rewriting with the initial checkpoint M_0, which acts as the model's pre-adaptation refinement operator. Keeping M_0 fixed preserves a reference for refining an evolving proposal distribution, bringing a complementary model perspective into supervision construction. In this design, M_t determines which questions are explored, while M_0 adds variation through rewriting without an external teacher.

Given image x, proposal q^{prop}, and evolution instruction r, rewriting produces

$$q^{\text{rw}} \sim M_0(\cdot \mid x, q^{\text{prop}}, r). \tag{3}$$

Instruction r targets greater visual inspection difficulty, evidence grounding, contextual reasoning, or spatial reasoning. Rewriting prompts also use model-extracted image facts as grounding evidence and difficult examples drawn from a bank of previously generated questions as refinement references. The complete prompts and their conditioning inputs are provided in Appendix C. The base model answers each rewritten question to form the candidate pool:

$$\mathcal{C}_t^{\text{rw}} = \{(x_i, q_{i,j}^{\text{rw}}, a_{i,j}^{\text{rw}})\}_{i,j}. \tag{4}$$

3.3 Question Filtering

Rewriting produces candidates with different levels of grounding and difficulty. The base model compares each rewritten triple $(x, q^{\text{rw}}, a^{\text{rw}})$ with its proposal $(x, q^{\text{prop}}, a^{\text{prop}})$, using the original question as the reference for improvement on the same image. It retains visually answerable, grounded candidates that improve perception or reasoning difficulty, discarding ambiguous, weakly grounded, unanswerable, or non-improving pairs. Filtering thus selects which refinements become training targets. The retained candidates form the supervision pool $\mathcal{D}^{(t)}$.

3.4 Training and Iterative Self-Improvement

3.4.1 Dual-Format Supervision

After the filtering process, the retained question-answer pairs are used to adapt the VLM. Our training stage mixes two supervision formats: QG predicts a question-answer pair from the image, while QA predicts an answer from the image and question. QG-format training targets questioning capability by teaching the model to generate both a question and its answer. QA-format training anchors the model's response to a supplied question, helping preserve answering ability during QG adaptation. The retained pairs $\mathcal{D}^{(t)} = \{(x_i, q_i, a_i)\}_{i=1}^N$ define two formats:

$$\mathcal{D}_{\text{QG}}^{(t)} = \{x_i \to (q_i, a_i)\}_{i=1}^N, \quad \mathcal{D}_{\text{QA}}^{(t)} = \{(x_i, q_i) \to a_i\}_{i=1}^N. \tag{5}$$

The training mixture $\mathcal{D}_{\text{train}}^{(t)}$ samples examples from $\mathcal{D}_{\text{QG}}^{(t)} \cup \mathcal{D}_{\text{QA}}^{(t)}$ under a fixed total budget. We train the model with standard supervised fine-tuning on the target tokens in both formats. This improves question generation while maintaining the model's answering ability.

3.4.2 ITERATIVE SELF-IMPROVEMENT

Fine-tuning M_t on $\mathcal{D}_{\text{train}}^{(t)}$ yields M_{t+1}. Initially, M_0 proposes, rewrites, and filters the data used to train M_1. In later rounds, the adapted checkpoint proposes new questions, while M_0 remains the rewriter and filter. The resulting supervision trains the next checkpoint, completing a self-improvement loop $M_t \to \mathcal{D}_{\text{train}}^{(t)} \to M_{t+1}$.

Overall, our method improves QG through the interaction of rewriting, filtering, and dual-format training. Rewriting expands the candidate pool toward harder and more visual-centric questions. Filtering removes invalid or non-improving generations, and dual-format supervision improves question generation while anchoring answering behavior. Together, these components enable the model to improve visual question generation while maintaining competitive downstream QA performance.

4 EXPERIMENTS

In this section, we evaluate whether our self-evolving visual questioner framework improves question generation (QG) quality while preserving question answering (QA) performance.

Our main experiments address four questions:

- Can our self-evolving framework improve QG quality within a single round?
- Does iterative self-improvement further improve QG quality across rounds?
- How do the effects vary across different generations and training settings?
- Does improved question quality provide more effective supervision for downstream QA training?

4.1 EXPERIMENTAL SETUP

We evaluate our self-evolving questioner framework on three VLM backbones: Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), and Qwen3VL-4B-Instruct (Bai et al., 2025a). Implementation details and training hyperparameters are provided in Appendix D.

We report results at up to four stages. **Base** denotes the original pretrained VLM. **First-round** denotes the model adapted with supervision constructed in the first round. **Second-round** denotes the model adapted with supervision constructed by reusing the first-round model as the question proposer. The **Third-round** checkpoints reuse the second-round models as proposers. Unless otherwise specified, training uses 10K examples and the QA+QG format described in Section 3.4.

To evaluate QG ability, models generate questions using shared prompts and decoding parameters. We then apply the visual questioning capability evaluation protocol described in Section 2.2. For the four individual-question dimensions, we use GPT-5-mini as an image-conditioned judge. The judge scores each image–question pair independently using the same rubric. Each dimension is scored on a 0–5 rubric and normalized to $[0, 1]$. Questioning diversity is computed within the generated question set for each image using Qwen3-Embedding-4B (Zhang et al., 2025b).

For Table 1, all five QG metrics use the same 100 held-out CVBench images, with 5 questions per image (500 questions per checkpoint). The self-evolution data are constructed from the SAT image collection, and no QG or QA evaluation image is used for training. We report five dimensions: Visual Search Difficulty (**Search**), Visual Evidence Coverage (**Coverage**), Visual Context Reasoning (**Context**), Visual Spatial Reasoning (**Spatial**), and Questioning Diversity (**Diversity**). To evaluate downstream answering ability, we use four benchmarks: CVBench (Tong et al., 2024), SparBench (Zhang et al., 2026), VStarBench (Wu & Xie, 2024), and RealWorldQA (xAI, 2024). CVBench is evaluated separately on its 2D and 3D subsets; QA Avg. is the unweighted mean of the resulting five accuracy scores.[2]

Backbone	Training stage	QG Dimension Scores					QG Avg.	QA Avg.
		Search	Coverage	Context	Spatial	Diversity		
Qwen2.5-VL-3B	Base	0.26	0.39	0.30	0.13	0.28	0.27	60.49
	First-round	0.43	0.55	0.55	0.37	0.35	0.45	62.26
	Second-round	**0.49**	**0.61**	**0.64**	0.40	**0.37**	0.50	**62.56**
	Third-round	0.48	**0.61**	**0.64**	**0.46**	**0.37**	**0.51**	61.87
Qwen2.5-VL-7B	Base	0.25	0.35	0.27	0.11	0.32	0.26	65.90
	First-round	0.46	0.57	0.59	0.35	0.38	0.47	**67.02**
	Second-round	**0.51**	0.59	**0.64**	0.40	**0.40**	**0.51**	66.50
	Third-round	**0.51**	**0.63**	**0.64**	**0.43**	0.35	**0.51**	66.80
Qwen3VL-4B	Base	0.38	0.51	0.45	0.18	0.26	0.36	**71.31**
	First-round	0.52	0.63	0.62	0.54	0.38	0.54	71.12
	Second-round	**0.56**	0.64	**0.65**	0.57	0.43	0.57	70.44
	Third-round	0.55	**0.66**	**0.65**	**0.58**	**0.45**	**0.58**	70.76

Table 1: **QG and QA performance across self-evolution rounds.** Across three backbones, self-evolution improves QG average over Base while largely preserving QA. QG scores lie in $[0, 1]$; QA Avg. averages five scores from four benchmarks, and bold marks per-backbone maximum value.

4.2 Main Results

Table 1 assesses overall QG improvement and QA preservation through comparisons with the base models, and characterizes continued self-evolution through comparisons between successive rounds.

Our framework improves questioning while preserving answering ability. Table 1 shows improvements over the base models in all four rubric dimensions after two rounds, with QG average gains of 0.21–0.25. The gains extend across the Qwen2.5-VL backbones and the stronger Qwen3VL-4B baseline, showing that refined self-supervision benefits questioners with different initial capabilities. The improved questions demand richer visual inspection and contextual and spatial reasoning, with higher question-set diversity across all three backbones. Because evaluation directly assesses unmodified model outputs, these results reflect the learned questioner's ability to ask better visual questions. These QG improvements are accompanied by largely preserved QA performance. Across rounds, both Qwen2.5-VL backbones keep QA averages above their baselines, and Qwen3VL-4B stays within 0.9% of its Base average. Together, these results show that self-evolution improves visual questioning while retaining answering ability on held-out benchmarks.

Iterative self-improvement refines questioning across rounds. QG averages exceed Base across all rounds. For 7B, they rise from 0.47 to 0.51 and remain at 0.51 in the third round; later 3B and 4B checkpoints have higher spatial-reasoning scores than First-round, and 3B diversity rises from 0.35 to 0.37 before remaining at 0.37. Across rounds, the adapted model supplies new proposals for rewriting and filtering, updating the supervision drawn from the same image pool. The observed trends support a self-improvement loop in which learning to ask questions also changes the questions available for subsequent training. Appendix G.1 reports the evaluation-sample variability used to interpret these between-round trends.

Comparison with other visual question generators. We compare our 3B questioners with LOVA3-7B (Zhao et al., 2024), a LLaVA-based model trained for visual question answering, asking, and assessment, and GPT-5.4 (Table 2). Direct prompting supplies GPT-5.4 with an image and our QG prompts to generate a question–answer pair without task-specific fine-tuning, rewriting, or filtering. Under our QG protocol, our third-round model leads in all five dimensions, averaging 0.17–0.18 higher without either model as a teacher. Appendix A provides prompt details and full scores.

Questioner	QG Avg.
LOVA3 (7B)	0.33
GPT-5.4 (prompted)	0.34
Ours (3B, first round)	0.45
Ours (3B, second round)	0.50
Ours (3B, third round)	**0.51**

Table 2: **QG comparison with LOVA3-7B and GPT-5.4.** Our 3B model leads on average across five QG dimensions in each round.

[2]Detailed accuracies are provided in Appendix Table 11.

Backbone	Setting	QG Dimension Scores					QG Avg.	QA Avg.
		Search	Coverage	Context	Spatial	Diversity		
Qwen2.5-VL-3B	QA-only	0.37	0.48	0.37	0.25	**0.37**	0.37	**62.40**
	QG-only	0.39	0.53	0.43	0.35	0.35	0.41	60.52
	QA+QG	**0.43**	**0.55**	**0.55**	**0.37**	0.35	**0.45**	62.26
Qwen2.5-VL-7B	QA-only	0.45	0.48	0.38	0.28	0.35	0.39	66.57
	QG-only	**0.48**	**0.57**	0.56	0.33	0.36	0.46	65.83
	QA+QG	0.46	**0.57**	**0.59**	**0.35**	**0.38**	**0.47**	**67.02**
Qwen3VL-4B	QA-only	0.47	0.57	0.47	0.37	0.33	0.44	70.38
	QG-only	0.51	0.61	0.59	**0.54**	0.35	0.52	70.12
	QA+QG	**0.52**	**0.63**	**0.62**	**0.54**	**0.38**	**0.54**	**71.12**

Table 3: **Effect of QA/QG supervision format.** We compare QA-only, QG-only, and QA+QG training across three backbones. QA+QG achieves the highest QG average on every backbone and higher QA accuracy than QG-only, supporting stronger questioning with retained answering ability.

Generation Process		QG Dimension Scores					QG Avg.
Rewriting	Filtering	Search	Coverage	Context	Spatial	Diversity	
–	–	0.26	0.39	0.30	0.13	0.28	0.27
✗	✗	0.31	0.50	0.47	0.20	0.26	0.35
✓	✗	0.35	0.52	0.51	0.29	0.32	0.40
✓	✓	**0.43**	**0.55**	**0.55**	**0.37**	**0.35**	**0.45**

Table 4: **Contribution of rewriting and filtering on Qwen2.5-VL-3B.** Starting from directly generated QA pairs, we add rewriting and then filtering; "–" denotes the unadapted base model. Each step raises QG average; the full pipeline leads on all five dimensions.

4.3 Ablation Studies

We next ablate supervision format, supervision type, and filtering to identify what drives these gains.

4.3.1 Effect of Supervision Format

Table 3 studies the effect of supervision format while keeping the retained self-generated examples fixed. We compare QA-only training, QG-only training, and the proposed QA+QG dual-format training. QA-only trains the model to answer generated questions, QG-only trains the model to generate question-answer pairs from the image, and QA+QG combines both formats.

QG-only improves the QG average over QA-only across all three backbones, showing the value of explicitly training image-to-question–answer generation. QA+QG achieves the highest contextual-reasoning score across all three backbones and the highest QG average on each backbone, while also attaining higher QA accuracy than QG-only. These results explain the role of dual-format supervision in the framework: QG-format examples improve questioning, and QA-format examples help preserve answering performance during adaptation.

4.3.2 Effect of Rewriting and Filtering

Table 4 ablates the effects of rewriting and filtering in Qwen2.5-VL-3B setting. Starting from direct self-generated supervision, we test whether refining candidate questions strengthens the subsequent QG training signal. Adding rewriting strengthens all five QG dimensions, expanding the pool toward harder, visually grounded questions.

Adding filtering yields the strongest performance across all five QG dimensions. The gains in visual search and spatial reasoning show that selecting refined supervision teaches the model to ask questions requiring closer visual inspection and richer reasoning. Rewriting and filtering thus play

Round-2 rewriter	Search	Coverage	Context	Spatial	Diversity	QG Avg.
Current model (M_1)	0.48	0.58	0.35	**0.43**	0.28	0.42
Frozen base (M_0)	**0.49**	**0.61**	**0.64**	0.40	**0.37**	**0.50**

Table 5: **Effect of the Round-2 rewriting checkpoint.** We compare the initial model M_0 and first-round model M_1 as rewriters for Qwen2.5-VL-3B. Frozen-base rewriting yields the higher QG average, with gains in four of five dimensions.

Data Source	QG Dimension Scores					QG Avg.	QA Avg.
	Search	Coverage	Context	Spatial	Diversity		
Base	0.26	0.39	0.30	0.13	0.28	0.27	60.49
Original SAT	**0.45**	0.49	0.21	0.22	0.17	0.31	57.46
Original SAT + Self-Supervision	0.44	0.54	0.54	0.32	0.34	0.44	61.28
Self-Supervision	0.43	**0.55**	**0.55**	**0.37**	**0.35**	**0.45**	**62.26**

Table 6: **Effect of the supervision source on Qwen2.5-VL-3B.** The three trained variants use SAT images, QA+QG supervision, and 10K examples. Rewritten and filtered self-supervision achieves higher QG and QA averages than original SAT annotations or their mixture.

complementary roles: rewriting expands the candidate pool, while filtering selects grounded, challenging supervision for training.

4.3.3 Effect of Rewriting Checkpoint

Table 5 compares the initial and First-round checkpoints as rewriters for constructing Round-2 supervision. Frozen-base rewriting increases visual search, evidence coverage, contextual reasoning, and diversity scores over the Round-1 rewriter, while spatial reasoning decreases from 0.43 to 0.40. These findings support complementary roles for the two checkpoints: the adapted proposer changes which questions are explored, while the initial model contributes its pre-adaptation refinement perspective.

4.3.4 Effect of Self-Supervision Data

Table 6 compares supervision sources using the same SAT images, QA+QG format, and 10K training budget. Original SAT uses directly sampled annotations; our self-supervision uses rewriting and filtering. At this matched budget, self-supervision achieves the highest QG and QA averages among the three training-data variants, including the mixture of source annotations and self-supervision. Relative to Original SAT, QG average rises by 0.14 and QA accuracy by 4.80%; the QG gains span evidence coverage, contextual and spatial reasoning, and diversity. With the image pool and training budget fixed, these results support refined self-supervision as a targeted learning signal for improving QG while preserving answering ability.

4.4 Do Better Questions Help Downstream Supervision?

To test whether improved QG yields more useful training data, we generate questions on the same SAT images using base Qwen2.5-VL-3B (Base-Q) or its first-round self-evolved checkpoint (Improved-Q), with GPT-5.4 answering in both settings. Each resulting 10K dataset trains a separate answerer initialized from Qwen2.5-VL-3B-Instruct using QA-only supervision. Both rows report the resulting answerers' QA accuracy after training.

In Table 7, the answerer trained on Improved-Q achieves a QA average 1.42% higher than the one trained on Base-Q, with gains on four of the five evaluation sets and the strongest improvement on CVBench-3D. These results connect improved question generation to useful downstream QA supervision.

Question source	CVB-2D	CVB-3D	SparBench	VStarBench	RWQA	Avg.
Base-Q	69.16	69.25	**30.21**	75.92	64.97	61.90
Improved-Q	**69.19**	**75.58**	29.78	**76.44**	**65.62**	**63.32**

Table 7: **Downstream QA training with base and improved questions.** Qwen2.5-VL-3B answerers train on 10K pairs with GPT-5.4 answers and questions from the base model (**Base-Q**) or its first-round checkpoint (**Improved-Q**). Improved-Q attains a QA average 1.42% higher, with gains on four of five evaluation sets.

5 Related Work

Visual Question Generation. VQG aims to generate natural-language questions grounded in images. Existing VQG methods are commonly built from external supervision, including VQA-style data (Mi et al., 2024; Shen et al., 2024; Vedd et al., 2022; Zhao et al., 2024), LLM-generated questions (Suwono et al., 2023; Zhang et al., 2025a), and controllable VQG data with predefined constraints such as answers, regions, knowledge, or difficulty levels (Fang et al., 2024; Mi et al., 2024; Uehara & Harada, 2023; Shen et al., 2024; Xie et al., 2025a;b). These methods establish the value of controlled, grounded question generation. We adapt the questioner through successive rounds of its own supervision, using the adapted model to propose the next round's questions.

VQG evaluation is also often tied to reference-based metrics such as BLEU, METEOR, CIDEr to compare similarity to given ground-truth questions (Mi et al., 2024; Vedd et al., 2022; Xie et al., 2025a;b). Recent work uses LLMs or MLLMs as judges for multimodal outputs or generated questions (Chen et al., 2024; Noorbakhsh et al., 2025; Yao et al., 2025), but mainly evaluates general multimodal responses or text-domain questions rather than fine-grained visual questioning.

Post-Training for VLMs. Post-training improves language and vision-language models with instruction data collected from humans, curated datasets, or stronger models (Liu et al., 2023; Wang et al., 2024; Zhou et al., 2026). However, constructing high-quality supervision is costly and difficult to scale. Recent self-improvement methods reduce this dependence by letting models generate, filter, critique, or judge their own training signals (Yuan et al., 2024; Shen et al., 2025; Deng et al., 2024; Cheng et al., 2025). However, these works mainly use self-generated supervision to improve answering-oriented capabilities, such as visual instruction following, reasoning, and QA performance (Wang et al., 2024; Deng et al., 2024; Cheng et al., 2025; Pan et al., 2026). VIGC, for example, generates and corrects visual instruction data for subsequent model tuning (Wang et al., 2024). Our framework makes the visual questioner itself the target of self-improvement while preserving answering ability. The adapted model proposes questions for refinement and dual-format training, closing the loop from learning to ask to constructing next-round supervision.

6 Limitations

Our evaluation covers rubric-based perception and reasoning scores and embedding-based diversity, with human annotations assessing agreement on the four rubric dimensions. The observed differences across evolution rounds motivate future work on adaptive stopping and selective supervision updates. Data construction adds generation and filtering cost beyond supervised fine-tuning alone; batching and lightweight refinement are promising ways to reduce this overhead.

7 Conclusion

We present a self-evolving visual questioner that improves QG through its own supervision while preserving answering ability. Proposal, rewriting, and filtering construct diverse, visually grounded training data, while dual-format training supports both questioning and answering. The adapted model then supplies proposals for the next round. Across three backbones, the framework delivers substantial QG gains. Ablations and downstream QA training demonstrate the value of refined self-supervision. These findings support self-evolution as a practical route to stronger visual questioning without an external teacher in the core learning loop.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

Kanzhi Cheng, Li YanTao, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. Vision-language models can self-improve reasoning via reflection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8876–8892, 2025.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems*, 37:131369–131397, 2024.

Wenhao Fang, Jiayuan Xie, Hongfei Liu, Jiali Chen, and Yi Cai. Diverse visual question generation based on multiple objects selection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(6):1–22, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Li Mi, Syrielle Montariol, Javiera Castillo Navarro, Xianjie Dai, Antoine Bosselut, and Devis Tuia. Convqg: Contrastive visual question generation with multimodal guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4207–4215, 2024.

Kimia Noorbakhsh, Joseph Chandler, Pantea Karimi, Mohammad Alizadeh, and Hari Balakrishnan. Savaal: Scalable concept-driven question generation to enhance human learning. *arXiv preprint arXiv:2502.12477*, 2025.

Zhiyu Pan, Yizheng Wu, Jiashen Hua, Junyi Feng, Shaotian Yan, Bing Deng, Zhiguo Cao, and Jieping Ye. Through the lens of contrast: Self-improving visual reasoning in vlms. *arXiv preprint arXiv:2603.02556*, 2026.

Kai Shen, Lingfei Wu, Siliang Tang, Fangli Xu, Bo Long, Yueting Zhuang, and Jian Pei. Ask questions with double hints: Visual question generation with answer-awareness and region-reference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9648–9660, 2024.

Zhanming Shen, Hao Chen, Yulei Tang, Shaolin Zhu, Wentao Ye, Xiaomeng Hu, Haobo Wang, Gang Chen, and Junbo Zhao. Cycle-instruct: Fully seed-free instruction tuning via dual self-training and cycle consistency. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5123–5137, 2025.

Nicholas Suwono, Justin Chen, Tun Hung, Ting-Hao Huang, I-Bin Liao, Yung-Hui Li, Lun-Wei Ku, and Shao-Hua Sun. Location-aware visual question generation with lightweight models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1415–1432, 2023.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Kohei Uehara and Tatsuya Harada. K-vqg: Knowledge-aware visual question generation for common-sense acquisition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4401–4409, 2023.

Nihir Vedd, Zixu Wang, Marek Rei, Yishu Miao, and Lucia Specia. Guiding visual question generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1640–1654, 2022.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5309–5317, 2024.

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*, pp. 13084–13094, 2024.

xAI. Realworldqa dataset. `https://huggingface.co/datasets/xai-org/RealworldQA`, 2024. Accessed: 2026-09-24.

Jiayuan Xie, Mengqiu Cheng, Xinting Zhang, Yi Cai, Guimin Hu, Mengying Xie, and Qing Li. Explicitly guided difficulty-controllable visual question generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25552–25560, 2025a.

Jiayuan Xie, Jiasheng Zheng, Wenhao Fang, Yi Cai, and Qing Li. Explicitly diverse visual question generation. *Neural Networks*, 184:107002, 2025b.

Zonghai Yao, Aditya Parashar, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, Zhichao Yang, and Hong Yu. Mcqg-srefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10728–10777, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

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, 2026.

Xinyu Zhang, Lingling Zhang, Yanrui Wu, Muye Huang, Wenjun Wu, Bo Li, Shaowei Wang, Basura Fernando, and Jun Liu. Diagram-driven course questions generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5995–6010, 2025a.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025b.

Henry Hengyuan Zhao, Pan Zhou, Difei Gao, Zechen Bai, and Mike Zheng Shou. Lova3: Learning to visual question answering, asking and assessment. *Advances in Neural Information Processing Systems*, 37:115146–115175, 2024.

Zhihan Zhou, Feng Hong, Jiaan Luo, Yushi Ye, Jiangchao Yao, Dongsheng Li, Bo Han, Ya Zhang, and Yanfeng Wang. Learning to instruct for visual instruction tuning. *Advances in Neural Information Processing Systems*, 38:47382–47409, 2026.

Model	Search	Coverage	Context	Spatial	Diversity	QG Avg.
LOVA3 (7B)	0.37	0.44	0.26	0.20	0.36	0.33
GPT-5.4 (direct prompting)	0.35	0.48	0.20	0.35	0.33	0.34
Ours, First-round	0.43	0.55	0.55	0.37	0.35	0.45
Ours, Second-round	**0.49**	**0.61**	**0.64**	0.40	**0.37**	0.50
Ours, Third-round	0.48	**0.61**	**0.64**	**0.46**	**0.37**	**0.51**

Table 8: **Per-dimension QG comparison with LOVA3-7B and GPT-5.4.** We evaluate Qwen2.5-VL-3B questioners after each self-evolution round; neither reference model provides training supervision. Our models attain higher QG averages in all three rounds, with the second and third rounds exceeding both references in every dimension.

A Comparison with Other Visual Question Generators

Comparison setup. We use the released LOVA3-LLaVA-v1.5-7B checkpoint (Zhao et al., 2024), whose training combines visual question answering, question generation, and question–answer assessment. GPT-5.4 is used as an image-conditioned question generator through prompting, with no task-specific fine-tuning. For both models, we use the same prompt families as for our questioners, producing 5 question–answer pairs per image. The prompts target local visual details, relations between visible elements, and less obvious image evidence, encouraging different question forms, image regions, and evidence types. Each prompt requests one image-grounded question–answer pair in a fixed output format. We extract the generated questions and evaluate them with our QG protocol, without rewriting or filtering them. For LOVA3, GPT-5.4, and our questioners, all five QG metrics use 100 images with 5 questions per image (500 questions per model or checkpoint). Diversity averages the 10 pairwise cosine distances within each image's five-question set. Our three checkpoint rows reuse the estimates in Table 1.

Our gains hold against other visual question generators. Table 8 compares our Qwen2.5-VL-3B questioners with LOVA3-7B and GPT-5.4 direct prompting under the same evaluation protocol. Our first-round checkpoint outperforms both references on all four difficulty dimensions. All three rounds attain higher QG averages than both references, and the second- and third-round checkpoints exceed both in every reported dimension. These comparisons contextualize performance against teacher-assisted or stronger-model alternatives; neither comparison model is used to construct or train our self-evolving checkpoints.

B Detailed Explanation of Questioning Capability Evaluation

Our evaluation protocol measures whether generated questions are visually grounded, challenging, and non-redundant. As summarized in Table 9, we evaluate generated questions at two levels. At the individual-question level, we measure four dimensions grouped into **perception difficulty** and **reasoning difficulty**. Perception difficulty captures what visual evidence must be found, while reasoning difficulty captures how that evidence must be interpreted. At the question-set level, we measure **questioning diversity**, which captures whether questions generated for the same image provide complementary rather than redundant supervision.

B.1 Perception Difficulty

Perception difficulty evaluates what visual evidence must be observed to answer a question. It captures the demand placed on visual inspection before reasoning: whether the answer depends on salient evidence, subtle details, localized regions, or broader image content. We decompose perception difficulty into two dimensions: *Visual Search Difficulty* and *Visual Evidence Coverage*.

Visual Search Difficulty. *Visual Search Difficulty* measures how difficult it is to locate and inspect the evidence needed to answer a question. Questions that depend on salient objects, obvious colors, or clearly visible main subjects receive lower scores. Questions receive higher scores when answering them requires careful inspection of subtle evidence, small or partially occluded objects, peripheral regions, fine-grained details, relative size or area, or counting.

Level	Dimension	What it measures
Question	**Visual Search Difficulty**	How difficult it is to locate the visual evidence needed to answer the question.
Question	**Visual Evidence Coverage**	How broadly the question depends on evidence across objects, regions, attributes, or scene cues.
Question	**Visual Context Reasoning**	Whether the question requires contextual interpretation of visible cues beyond direct recognition.
Question	**Visual Spatial Reasoning**	Whether the question requires reasoning over spatial relations, layout, orientation, or scene structure.
Set	**Questioning Diversity**	How non-redundant the questions generated for the same image are.

Table 9: Summary of our visual questioning capability evaluation dimensions.

Visual Evidence Coverage. *Visual Evidence Coverage* measures how broadly a question depends on evidence across the image. Questions centered on a single dominant object or a single simple attribute receive lower scores. Questions receive higher scores when they require evidence from multiple objects, regions, attributes, scene cues, or the overall arrangement of the image. This dimension is independent of search difficulty: a question may rely on broad image evidence even if each individual cue is easy to locate.

B.2 Reasoning Difficulty

Reasoning difficulty evaluates how the observed visual evidence must be used to answer a question. While perception difficulty focuses on finding and inspecting evidence, reasoning difficulty focuses on the inference required after the relevant evidence has been identified. We decompose reasoning difficulty into two dimensions: *Visual Context Reasoning* and *Visual Spatial Reasoning*.

Visual Context Reasoning. *Visual Context Reasoning* measures whether a question requires interpreting what visible cues imply beyond direct recognition or description. Questions that can be answered by simply naming objects, reading visible attributes, or describing directly observable content receive lower scores. Questions receive higher scores when they require inferring the scene context, object state, activity, function, intention, cause, or likely situation suggested by the visual evidence.

Visual Spatial Reasoning. *Visual Spatial Reasoning* measures whether a question requires reasoning about spatial relationships among visible elements. Questions that do not depend on spatial structure receive lower scores. Questions receive higher scores when they require comparing positions, orientations, distances, containment, occlusion, accessibility, layout, object interactions, or multi-hop spatial relations within the scene.

B.3 Questioning Diversity

Questioning Diversity is measured at the question-set level because diversity cannot be determined from a single question in isolation. For each image x, we consider the generated question set $\mathcal{Q}_x = \{q_1, \ldots, q_m\}$. For the main results and their bootstrap analysis, we use five questions ($n = 5$) on each of the same 100 images (500 questions per checkpoint). Other sampled-set analyses average five draws of $n = 5$ questions per image. Each question is encoded by Qwen3-Embedding-4B with its last-token pooling and ℓ_2-normalized:

$$z_i = \frac{f(q_i)}{\|f(q_i)\|_2}. \tag{6}$$

We then compute the average pairwise cosine similarity within each question set:

$$s_{\mathrm{emb}}(\mathcal{Q}_x) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} z_i^\top z_j. \tag{7}$$

The diversity score is defined as the corresponding average pairwise cosine distance:

$$d_{\mathrm{div}}(\mathcal{Q}_x) = 1 - s_{\mathrm{emb}}(\mathcal{Q}_x). \tag{8}$$

Question Proposal

```
List n1 topics that you can answer questions about. Choose a topic
    uniformly from this list, and state it. Then write 60 subtopics
    about the chosen topic. Then choose a subtopic uniformly from
    this list, and state it. Then write a question that is not about
     the subtopic, but can only be answered with expertise in the
    subtopic. Then write the answer. Both the question and answer
    should be long. The name of the subtopic should not appear in
    the question, and none of the words in the subtopic should be
    reused in the question.
The question must rely on visual information such that its answer
    would change or be impossible to determine without the image.
    The answer should be deterministic and answerable with a few
    sentences. You should generate only ONE focused question, not
    multiple sub-questions. IMPORTANT: Do NOT create compound or
    multi-part questions by joining sub-questions with 'and' or '
    what'. A question like 'How does X affect Y, and what in the
    image shows Z?' is FORBIDDEN. It is two easy questions
    concatenated, not a harder one. Instead, make the SINGLE
    question harder by requiring deeper reasoning: multi-hop
    inference, counterfactual/causal reasoning, cross-region
    comparison, or quantitative/ordinal judgment. The answer must be
     DIFFERENT from the original answer.
Return the result in JSON format: {'Question': 'xxx', 'Thinkings':
    'Your selection process'}.
```

Figure 2: Question proposal prompt used to produce the initial question-answer pair.

Higher values indicate that the questions generated for the same image are semantically less redundant and cover more distinct aspects of the image. Lower values indicate repeated templates, near-paraphrases, or duplicated answer targets. This set-level metric complements the individual-question rubric scores by measuring whether the generated supervision is diverse across questions, rather than only strong for each question independently.

C Detailed Prompts

In this section, we provide the detailed prompts used for our data generation process and evaluation process. Table 10 summarizes the prompts used in our framework. The full prompt texts are shown in Fig. 2–Fig. 8.

Stage	Reference
Question Proposal	Fig. 2
Rewriting with visual facts	Fig. 3
Rewriting with hard examples	Fig. 4
Filtering	Fig. 6 & 7
QG evaluation	Fig. 8

Table 10: Overview of prompts used in the self-evolving visual questioner framework and visual questioning capability evaluation.

D Implementation Details

Training setup. All models are trained with standard supervised fine-tuning from their corresponding instruction-tuned checkpoints. We conduct the main experiments with Qwen2.5-VL-3B-Instruct,

Rewriting Prompt for Broader Questions

```
You are a Question generator. Your objective is to rewrite a given
    Q&A into a genuinely harder version that requires more visual
    evidence or reasoning steps. Make it harder by increasing the
    DEPTH of reasoning required. Examples include evidence selection
    , object-role comparison, spatial relation reasoning, scene-
    function inference, and multi-object grounding. Do NOT make it
    harder by simply appending more sub-questions or extra clauses.
    Do NOT generate a question that asks the same thing as the
    original or is just a paraphrase. Do NOT start your rewritten
    question with the same first 5 words as the seed question. The
    question should not become broader, more subjective, or more
    generic. The added constraint must change the type or depth of
    reasoning required, not just expand the question's length.

Use the extracted visual facts below as grounding evidence. Only
    use facts that are clearly supported by the image.

Extracted visual facts:
{rag_facts}

Question: {seed_question}
Answer: {seed_answer}

Constraints:
- Require at least 2 concrete visible objects, regions, or
    relations.
- The question must contain a hidden intermediate inference.
- The intermediate inference step must not be given away in the
    question.
- Do NOT ask for subjective preference or broad commonsense.
- Do NOT merely paraphrase the seed question with synonyms.
- Ensure all generated data is consistent with the image content.
- The question must rely on visual information such that its answer
     would change or be impossible to determine without the image.
- Generate only ONE focused question, not multiple sub-questions.
- Make the SINGLE question harder by requiring deeper reasoning:
    multi-hop inference, counterfactual/causal reasoning, cross-
    region comparison, or quantitative/ordinal judgment.
- 'reasoning_steps' MUST contain at least 2 non-trivial, concrete
    operations, not just 'step 1' or 'step 2'.

Return the result in JSON format:
{'Rewritten question': 'one single syntactic question ending with
    ?', 'visible_evidence': ['at least two concrete visible objects/
    regions/relations required to answer, using the EXACT object
    names that appear in your question'], 'intermediate_inference':
    'the hidden bridging inference required before the final answer
    ', 'reasoning_steps': ['<concrete step, e.g. Locate the red
    chair>', '<concrete step, e.g. Compare its distance to the desk
    lamp vs the blue chair>'], 'why_harder_than_seed': 'why this
    requires more grounded reasoning than the seed',}
```

Figure 3: Rewriting prompt that rewrites a question proposal using extracted visual facts as grounding evidence.

Qwen2.5-VL-7B-Instruct, and Qwen3VL-4B-Instruct, and an additional transfer experiment with InternVL3.5-4B. Unless otherwise specified, each model is trained on 10,000 retained question-answer pairs produced by our self-evolution pipeline. We use a training batch size of 2 and AdamW

Rewriting Prompt for Harder Questions

```
You are a Question generator. Your objective is to rewrite a given
    Q&A into a genuinely harder version that requires more visual
    evidence or reasoning steps. Make it harder by increasing the
    DEPTH of reasoning required. Examples include evidence selection
    , object-role comparison, spatial relation reasoning, scene-
    function inference, and multi-object grounding. Do NOT make it
    harder by simply appending more sub-questions or extra clauses.
    Do NOT generate a question that asks the same thing as the
    original or is just a paraphrase. Do NOT start your rewritten
    question with the same first 5 words as the seed question. The
    question should not become broader, more subjective, or more
    generic. The added constraint must change the type or depth of
    reasoning required, not just expand the question's length.

The following questions were generated for similar images and are
    rated as HARD:
{examples_text}

Rewrite the following Q&A so it is harder than the examples while
    staying tightly grounded in this image. It must require more
    reasoning steps, finer perceptual discrimination, or a deeper
    multi-hop chain.

Question: {seed_question}
Answer: {seed_answer}

Examples of GOOD rewritten questions:
Example 1:
Q: 'Which chair, the wooden chair or the metal chair, is positioned
     closer to the desk lamp?'
visible_evidence: ['wooden chair', 'metal chair', 'desk lamp']
reasoning_steps: ['Locate the wooden chair and measure its distance
     to the desk lamp', 'Locate the metal chair and measure its
    distance to the desk lamp', 'Compare the two distances']
```

Figure 4: (Part 1) Hard-question rewriting prompt: instruction, seed question, and example. Continued in Fig. 5.

as the optimizer. The learning rate is set to 2×10^{-5} with a cosine scheduler and a warm-up ratio of 0.1. All models are trained for one epoch.

Downstream QA supervision study. For Section 4.4, Base-Q and Improved-Q denote question sources: the initial Qwen2.5-VL-3B-Instruct model and its first-round QA+QG checkpoint, respectively. Both answerers start from Qwen2.5-VL-3B-Instruct and train for one epoch on $10{,}000$ QA-only examples with matched SAT images and GPT-5.4 answers. We use LoRA rank 8, per-device batch size 2, gradient accumulation 2, a cosine schedule, and warm-up ratio 0.1. The learning rate is 5×10^{-6} for Base-Q and Improved-Q.

Computing infrastructure and budget. Training and evaluation use NVIDIA RTX A5000, RTX A6000, and A100 GPUs, depending on the backbone and available cluster resources. Most training, generation, and evaluation runs use a single GPU per run. For data generation, producing one training-ready set of $10{,}000$ self-supervision question-answer pairs requires approximately 24 GPU hours with 16-way parallel generation. For supervised fine-tuning, training one model on $10{,}000$ retained examples for one epoch takes approximately $1-3$ GPU hours, depending on the model size. For evaluation, running the full benchmark suite for one checkpoint takes approximately $4-7$ GPU hours, depending on the model size. Across the main experiments, ablations, and evaluation runs,

Rewriting Prompt for Harder Questions (Continued)

```
Constraints:
- Do not reuse the same reasoning pattern as any example above.
- The added difficulty must come from deeper reasoning, not longer
    question text.
- Require at least 2 concrete visible objects, regions, or
    relations.
- The question must contain a hidden intermediate inference.
- Do NOT merely paraphrase the seed question with synonyms.
- Ensure all generated data is consistent with the image content.
- The question must rely on visual information such that its answer
     would change or be impossible to determine without the image.
- Generate only ONE focused question, not multiple sub-questions.
- 'reasoning_steps' MUST contain at least 2 non-trivial, concrete
    operations, not just 'step 1' or 'step 2'.

Return the result in JSON format:
{'Rewritten question': 'one single syntactic question ending with
    ?', 'visible_evidence': ['at least two concrete visible objects/
    regions/relations required to answer, using the EXACT object
    names that appear in your question'], 'intermediate_inference':
    'the hidden bridging inference required before the final answer
    ', 'reasoning_steps': ['<concrete step, e.g. Locate the red
    chair>', '<concrete step, e.g. Compare its distance to the desk
    lamp vs the blue chair>'], 'why_harder_than_seed': 'why this
    requires more grounded reasoning than the seed', '
    final_answer_type': 'short phrase|entity|comparison|count|yes_no
    ', }
```

Figure 5: (Part 2) Hard-question rewriting prompt: constraints and output schema. Continued from Fig. 4.

the total computational budget is estimated from these per-run costs and includes both successful runs and exploratory runs used to tune training settings. All experiments are run on a Linux GPU cluster with standard CUDA/PyTorch-based training infrastructure.

Evaluation setup. QG generation uses a held-out image pool. The QG average combines the five dimension means using unrounded values. We use `gpt-5-mini` as the judge model for individual-question evaluation, including visual search difficulty, visual evidence coverage, visual context reasoning, and visual spatial reasoning. The `gpt-5-mini` judge is held out from data construction and training and is used only to evaluate raw model outputs after training. For set-level diversity evaluation, we use `Qwen/Qwen3-Embedding-4B` as the embedding model and compute semantic distances among questions generated for the same image. For QG evaluation, we compute scores by averaging evaluation results across generated questions or images, depending on the evaluation dimension. We estimate metric variability by bootstrap resampling the evaluation set and report the resulting bootstrap-estimated standard deviations in Appendix G.1. QA scoring details are provided in Appendix E.

Software and evaluation implementations. For model training and question generation, we use the official Qwen2.5-VL and Qwen3-VL implementations through the standard PyTorch and Hugging Face Transformers package. For diversity evaluation, we use `Qwen/Qwen3-Embedding-4B` as the embedding model and compute semantic distances among questions generated for the same image. For downstream QA evaluation, we use VLMEvalKit to evaluate benchmark performance and follow the standard evaluation protocols and answer-matching rules of each benchmark. We do not modify the underlying model architectures or external library implementations.

Filtering Prompt (Part 1)

```
You are a strict filtering judge for visual question generation.

Your task is to evaluate whether a generated question-answer pair
    should be kept for training. You must judge the question using
    the provided image, the seed question-answer pair, and the
    generated question-answer pair.

You must perform four checks:
1. Entity extraction: extract all noun or noun-phrase entities
    mentioned in the generated question.
2. Visual entity difficulty: categorize each shared visual entity
    as EASY or HARD based on the image.
3. Reasoning difficulty: infer the cognitive reasoning steps
    required to answer the generated question.
4. filter decision: decide whether the generated question is better
     than the seed question and should be kept.

Definitions:
- EASY visual entity: prominent, unobstructed, and immediately
    identifiable in the image.
- HARD visual entity: tiny, occluded, blurry, low-contrast,
    visually similar to nearby objects, peripheral, or requires
    careful inspection.
- Reasoning step: one cognitive operation required to answer the
    question, such as identifying objects, comparing quantities,
    interpreting spatial relations, inferring function, integrating
    multiple cues, or reasoning about cause/effect.
- Do not count meta steps such as "look at the image" or "analyze
    the question".
- Do not include verification or double-checking steps.
- Do not include the final answer as a reasoning step.

Keep the generated question only if ALL conditions below are
    satisfied:
- It is answerable from the image.
- Its answer is consistent with the image.
- It is not a trivial rephrasing of the seed question.
- It is visually grounded and depends on image evidence.
- It is at least as difficult as the seed question in one or more
    of the following dimensions:
  visual perception difficulty, text perception difficulty, or
      reasoning difficulty.
- It does not become vague, subjective, overly broad, or multi-
    question.
- It asks one focused question, not several questions joined
    together.
- Its answer is deterministic and concise.
```

Figure 6: (Part 1) Filtering prompt used to determine whether a rewritten question should be retained in the self-evolving visual questioner framework.

E Detailed QA Benchmark Results

Table 11 reports benchmark-level QA results for the main experiments and ablations.

Across the three self-evolution rounds, both Qwen2.5-VL backbones retain QA averages above their base models, with the clearest shared gains on CVBench-3D. At the third round, these gains are 5.50% and 4.50% for 3B and 7B, respectively. The gains are not uniform across benchmarks: for

Backbone	Training stage / Data source	CVB-2D	CVB-3D	SparBench	VStarBench	RWQA	Avg.
Qwen2.5-VL-3B	Base	66.75	65.50	28.89	74.39	66.92	60.49
	First-round	69.35	71.50	30.50	74.87	65.10	62.26
	Second-round	69.31	72.17	31.24	75.39	64.70	62.56
	Third-round	68.95	71.00	30.33	74.87	64.18	61.87
	QA-only	69.28	71.67	30.96	74.87	65.23	62.40
	QA+QG	69.35	71.50	30.50	74.87	65.10	62.26
	Original SAT	64.87	61.75	31.06	68.59	61.05	57.46
	Original SAT + Self-Supervision	68.74	69.67	30.92	73.30	63.79	61.28
	Self-Supervision	69.35	71.50	30.50	74.87	65.10	62.26
Qwen2.5-VL-7B	Base	75.67	74.83	33.81	76.44	68.76	65.90
	First-round	75.98	78.08	33.50	77.49	70.07	67.02
	Second-round	74.80	78.75	33.50	76.44	69.02	66.50
	Third-round	75.20	79.33	33.35	76.44	69.67	66.80
	QA-only	75.19	77.67	33.61	77.49	68.89	66.57
	QA+QG	75.98	78.08	33.50	77.49	70.07	67.02
Qwen3VL-4B	Base	78.83	91.92	36.16	78.53	71.11	71.31
	First-round	78.06	91.08	36.29	79.06	71.11	71.12
	Second-round	77.91	91.50	35.51	78.53	68.76	70.44
	Third-round	76.01	91.25	35.58	80.10	70.85	70.76
	QA-only	77.39	90.83	35.71	77.49	70.46	70.38
	QA+QG	78.06	91.08	36.29	79.06	71.11	71.12

Table 11: **Benchmark-level QA performance across rounds and supervision ablations.** We report CVBench-2D/3D, SparBench, VStarBench, and RealWorldQA accuracy with an unweighted average; First-round is the default QA+QG setting. Both Qwen2.5-VL backbones retain QA averages above Base across rounds; Self-Supervision has the highest average among 3B data sources.

3B, RealWorldQA remains below the base model in all three rounds, while the improvements on CVBench-2D/3D and SparBench more than offset this decrease in the average. Qwen3VL-4B, the strongest base model, stays within 0.9% of its base average across rounds, keeps CVBench-3D above 91%, and improves VStarBench in the first and third rounds (up to $+1.57\%$). These benchmark-level profiles complement the aggregate results in Table 1, showing how answering performance is retained alongside QG improvement.

QA+QG yields higher QA averages than QA-only for 7B and 4B and a similar average for 3B. Together with the QG improvements in Table 3, these comparisons support the dual-format objective of stronger questioning with retained answering. For 3B, Self-Supervision improves over Original SAT on four of five QA evaluation sets, with the largest gain on CVBench-3D (9.75%); the only exception is SparBench, where the two are within 0.6%. At a matched training budget, this comparison supports the usefulness of the self-generated and refined supervision.

F Human and QG Evaluation Alignment

F.1 Human Annotation Protocol

To validate the reliability of our GPT-based question-generation evaluation, we conduct a human–GPT alignment study on a randomly sampled set of 100 generated visual questions. Four annotators were recruited internally from student researchers familiar with vision-language tasks. Participation was voluntary and unpaid, and no private or sensitive information was collected from the annotators. The annotation task only involves scoring generated questions according to the same rubrics used by our automatic evaluator.

Annotators are shown the image, the generated question, and the scoring instructions for each dimension. They are asked to independently score four QG dimensions: visual search difficulty, visual evidence coverage, visual context reasoning, and visual spatial reasoning. The full annotation interface and instruction screenshots are shown in Figure 9.

The task does not ask annotators to provide personal information, identify individuals in the images, infer protected attributes, or make decisions about people. Annotators are instructed to focus only

QG Dimension	GPT–Human	Random baseline	Human–Human
Visual Search Difficulty	0.570	0.000 ± 0.089	0.467
Visual Evidence Coverage	0.527	0.001 ± 0.078	0.459
Visual Context Reasoning	0.637	0.000 ± 0.083	0.604
Visual Spatial Reasoning	0.473	0.001 ± 0.090	0.683

Table 12: **Human validation of the GPT-based QG evaluator.** Four annotators rate 100 questions; GPT–Human reports Spearman correlation with mean human scores, and Human–Human reports mean pairwise correlation. GPT–human correlations exceed the near-zero permutation baseline in all four dimensions, supporting the rubric-based evaluator.

Backbone	Training stage	Search	Coverage	Context	Spatial	Diversity
Qwen2.5-VL-3B	Base	0.26 ± 0.014	0.39 ± 0.028	0.30 ± 0.019	0.13 ± 0.012	0.28 ± 0.013
	First-round	0.43 ± 0.013	0.55 ± 0.014	0.55 ± 0.014	0.37 ± 0.014	0.35 ± 0.009
	Second-round	0.49 ± 0.012	0.61 ± 0.012	0.64 ± 0.014	0.40 ± 0.016	0.37 ± 0.011
Qwen2.5-VL-7B	Base	0.25 ± 0.015	0.35 ± 0.029	0.27 ± 0.021	0.11 ± 0.014	0.32 ± 0.011
	First-round	0.46 ± 0.016	0.57 ± 0.025	0.59 ± 0.013	0.35 ± 0.017	0.38 ± 0.011
	Second-round	0.51 ± 0.012	0.59 ± 0.011	0.64 ± 0.014	0.40 ± 0.015	0.40 ± 0.010
Qwen3VL-4B	Base	0.38 ± 0.018	0.51 ± 0.034	0.45 ± 0.026	0.18 ± 0.016	0.26 ± 0.014
	First-round	0.52 ± 0.015	0.63 ± 0.013	0.62 ± 0.008	0.54 ± 0.019	0.38 ± 0.009
	Second-round	0.56 ± 0.007	0.64 ± 0.009	0.65 ± 0.012	0.57 ± 0.012	0.43 ± 0.006

Table 13: **Sampling variability of multi-round QG results.** Entries report means $\pm$ image-level bootstrap standard deviations. After two rounds, all four rubric scores exceed Base across the three backbones; the standard deviations quantify the sampling variability of each dimension.

on whether the question is grounded in the image and how much visual perception or reasoning is required to answer it. Participation is voluntary, and no private or sensitive information is collected from the annotators. The sampled questions are generated from existing public research datasets and are screened to avoid clearly offensive or unsafe content before annotation. The collected annotations are used only to measure agreement with GPT-based evaluation and are reported only in aggregate form through correlation statistics.

F.2 Human Validation of the Evaluator

On 100 questions rated by four annotators, GPT–human Spearman correlations range from 0.473 to 0.637, versus 0.459–0.683 for mean pairwise human agreement. The observed GPT–human correlations provide human-annotation support for the rubric-based evaluation. Table 12 reports agreement for each dimension together with a permutation baseline.

GPT–Human correlations are consistently above the permutation baseline, which is centered near zero. The GPT–Human correlations are therefore broadly comparable to the level of agreement among independent annotators, although spatial reasoning retains a noticeable gap.

These results support the GPT-based judge as a scalable but imperfect proxy for human evaluation rather than a replacement for human judgment. The expanded study strengthens confidence in aggregate comparisons, while the moderate correlations motivate cautious interpretation of small metric differences and continued validation with human annotations.

G Additional Robustness and Generalization Analyses

G.1 Bootstrap Variability

Table 13 evaluates every checkpoint on the same 100 CVBench images and 5 generated questions per image (500 questions per checkpoint). For each dimension, we average within images and then across images. Diversity averages the 10 pairwise cosine distances among each image's 5 questions. These are sampling-uncertainty estimates, not the standard deviation of individual question scores.

Stage	Search	Coverage	Context	Spatial	Diversity	QG Avg.
Base	0.44	0.55	0.54	0.40	**0.35**	0.46
First-round	**0.46**	**0.60**	**0.64**	**0.46**	0.34	**0.50**

Table 14: **One-round self-evolution on InternVL3.5-4B.** Applying the pipeline to this non-Qwen backbone improves all four difficulty scores and the QG average.

Stage	Search	Coverage	Context	Spatial
Base	0.44	0.51	0.34	0.27
First-round	0.46	0.57	0.48	0.44
Second-round	**0.49**	**0.62**	**0.54**	**0.46**

Table 15: **Question-generation transfer to RealWorldQA images.** We evaluate the SAT-trained Qwen2.5-VL-3B checkpoints on this separate image pool without additional training. All four difficulty scores increase from Base through the second round.

For the supervision-format ablation in Table 3, we report point estimates without uncertainty estimates. QA+QG reuses the First-round results. The single-format rubric evaluations use 100 images; their diversity estimates use 100 images with 5 sampled five-question sets per image.

Across all three backbones, the Base-to-second-round gains in the four rubric dimensions are directionally consistent and generally large relative to the estimated variability. Diversity increases from Base to the second round across all three backbones, by 0.09, 0.08, and 0.17 for 3B, 7B, and 4B, respectively. The primary claim concerns improvement over the base model and does not depend on any single fine-grained comparison. Some smaller Round-1-to-Round-2 differences are comparable to sampling variability; we therefore interpret these secondary differences as trends rather than definitive advantages. These bootstrap estimates describe evaluation-sample variability, not variation across independent training seeds or repeated judge calls. Table 13 covers the main QG comparison; it does not establish significance for the ablations or for pairwise differences merely by comparing marginal standard deviations.

G.2 Transfer to a Non-Qwen Backbone

To test whether the framework transfers beyond the Qwen family, we apply one round of the same pipeline to InternVL3.5-4B.

The QG average improves from 0.46 to 0.50, with gains in all four difficulty dimensions and a small diversity change from 0.35 to 0.34. This provides preliminary evidence that the method is not specific to Qwen backbones.

G.3 Cross-Image-Pool Transfer

We evaluate the same Qwen2.5-VL-3B checkpoints on a separate RealWorldQA image pool without retraining or prompt modification.

All four dimensions improve from Base to Round 1 and again from Round 1 to Round 2. Because self-evolution uses the separate SAT image collection, these results provide evidence that the gains transfer across image domains rather than reflecting reuse of evaluation images.

G.4 Robustness of the Diversity Trend

We recompute questioning diversity with two additional embedding families on the 100-image generation pool, averaging five sampled five-question sets per image. Because cosine-distance scales differ across encoders, only within-column trends should be compared. All three encoders preserve the Base $<$ First-round $<$ Second-round trend, indicating that the diversity conclusion is not an artifact of a single embedding model.

Stage	Qwen3-Embedding-4B	bge-m3	gte-large-en-v1.5
Base	0.28	0.27	0.30
First-round	0.35	0.33	0.31
Second-round	**0.37**	**0.35**	**0.32**

Table 16: **Diversity trends across embedding models.** We compute question-set diversity with Qwen3-Embedding-4B, bge-m3, and gte-large-en-v1.5. All three encoders preserve the Base $<$ First-round $<$ Second-round ordering, supporting a consistent diversity gain across rounds.

H Qualitative Analysis

We provide qualitative examples in Fig 10 and 11 comparing questions generated by the base model M_0, the first-round model M_1, and the second-round model M_2 on the same images. These examples illustrate how our method gradually improves the model's QG capability. The base model typically produces simple recognition or attribute-based questions that are valid but require limited visual inspection or reasoning. After adaptation, the model generates questions that are more grounded in specific visual evidence and involve richer reasoning signals, including cross-region reflection reasoning, relative-size comparison, depth/layout understanding, and fine-grained counting with exclusion. The second-round model further improves question specificity and relation-centered reasoning, consistent with the quantitative gains observed in visual search difficulty, evidence coverage, visual context reasoning, spatial reasoning, and diversity.

I AI Use Statement

In this work, we used generative AI tools to improve the readability and linguistic clarity of the manuscript, review the code implementing our method, and modify the code used to generate figures.

We did not use generative AI tools to formulate or refine hypotheses, design or provide feedback on the research methodology or experiments, or develop theoretical models or conceptual frameworks. We also did not use them for translation, qualitative or thematic data analysis, interpretation of results, synthetic data generation, or dataset cleaning and reformatting. Tasks involving mathematical claims or formal proofs are not applicable to this work.

We reviewed all AI-assisted manuscript edits and code-review suggestions. We checked that readability edits preserved the intended meaning and examined each code-review suggestion against the corresponding implementation. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

Filtering Prompt (Part 2)

```
Inputs:
Image: [VISUAL_INPUT]

Seed question:
{seed_question}

Seed answer:
{seed_answer}

Generated question:
{generated_question}

Generated answer:
{generated_answer}

Output a SINGLE minified JSON object only.
Do not output markdown, explanations, comments, or extra text.
The JSON must start with { and end with }.

Required JSON schema:
{
  "entity_extraction": {
    "generated_question_entities": ["entity1", "entity2"]
  },
  "visual_entity_difficulty": {
    "easy_count": 0,
    "hard_count": 0,
    "categorized_entities": [
      {
        "entity": "name",
        "difficulty": "easy|hard",
        "reason": "short reason"
      }
    ]
  },
  "reasoning_difficulty": {
    "count": 0,
    "evidence": [
      "step1",
      "step2"
    ]
  }
}

Important rules:
- Analyze every extracted entity that is visible or potentially
    grounded in the image.
- The sum of easy_count and hard_count must equal the number of
    unique visual entities analyzed.
- If the generated question cannot be answered from the image, set
    decision to "discard".
- If the generated answer contradicts the image, set decision to "
    discard".
- If the question contains multiple sub-questions, set
    single_focused_question to false and decision to "discard".
- If none of improves_visual_perception, improves_text_perception,
    or improves_reasoning is true, set decision to "discard".
```

Figure 7: (Part 2) Filtering prompt: inputs, output schema, and decision rules.

Question Quality Evaluation Prompt

```
Use only the image and the current question. Ignore model identity
    and do not compare with other questions. Score what the question
     requires to answer, not how long or fluent it is. Give credit
    to visually grounded conceptual or functional questions when the
     image contains concrete evidence; penalize generic world-
    knowledge answers, unsupported assumptions, and unanswerable
    questions.

Record:
{record_json}

Score independently from 0 to 5:
- visual_search_difficulty: visual search/inspection effort. 0=no
    image needed; 1=obvious evidence; 3=specific region or nearby
    items; 5=multi-region search, counting, distance/area, or easy-
    to-miss fine detail.
- visual_evidence_coverage: breadth of concrete visual evidence. 0=
    no visible evidence; 1=one dominant object/scene; 3=two objects
    or one local region; 5=broad scene evidence, multiple important
    regions, or overall arrangement.
- visual_context_reasoning: conceptual interpretation beyond direct
     observation. 0=non-visual/unsupported; 1=direct identifying,
    reading, counting, locating, or comparing; 2=simple visible
    state/action/category judgment; 3=single-step interpretation
    from one cue; 4=multi-cue scene/situation interpretation; 5=non-
    obvious explanation or implication requiring weighed visual cues
    .
- visual_spatial_reasoning: use of layout/position/object relations
    . 0=none; 1=absolute location; 2=one simple relation; 3=one
    spatial comparison; 4=multiple linked relations; 5=complex
    spatial inference such as path, occlusion, containment,
    stability, or scene structure.

Return JSON only:
{
  "visual_search_difficulty": int,
  "visual_evidence_coverage": int,
  "visual_context_reasoning": int,
  "visual_spatial_reasoning": int
}
```

Figure 8: Question quality evaluation prompt used to score generated questions along visual search, evidence coverage, contextual reasoning, and spatial reasoning dimensions.

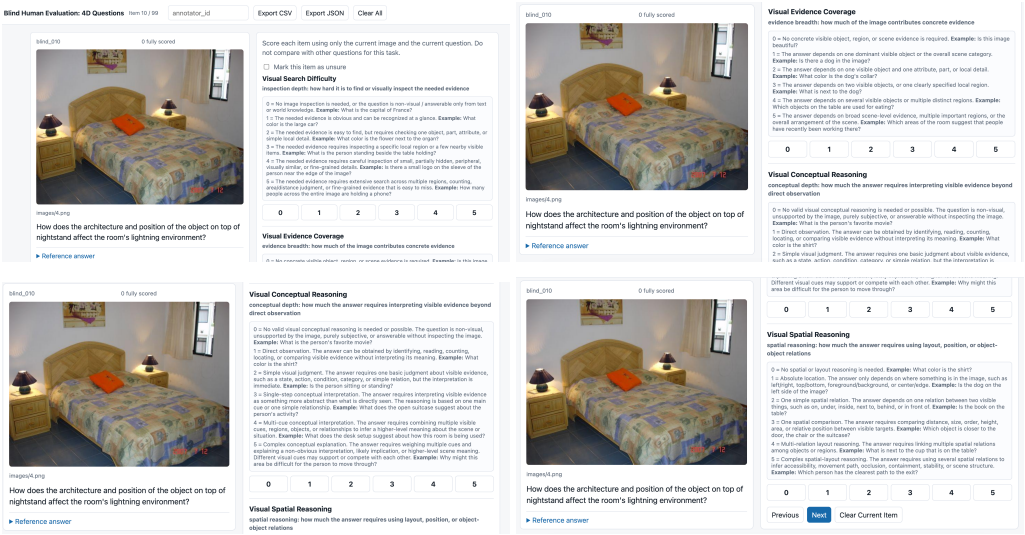

Figure 9: **Human annotation interface for evaluator validation.** Annotators view the image and question and independently score visual search, evidence coverage, contextual reasoning, and spatial reasoning. The shared rubrics provide a common basis for comparing human and automatic ratings.

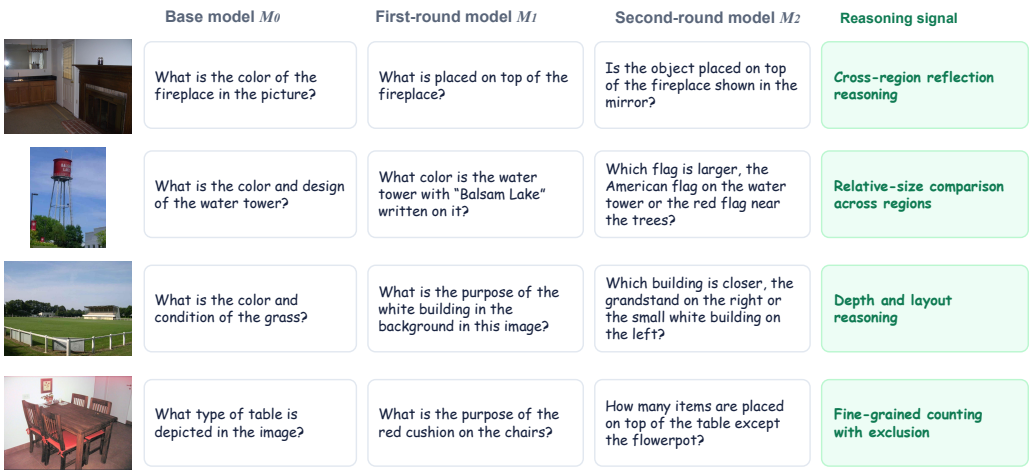

Figure 10: **Visual reasoning in questions generated across rounds.** We compare the base (M_0), first-round (M_1), and second-round (M_2) questioners on the same images. Second-round questions require reflection reasoning, relative-size comparisons, depth judgments, and counting with exclusion, illustrating richer visual inspection and spatial reasoning.

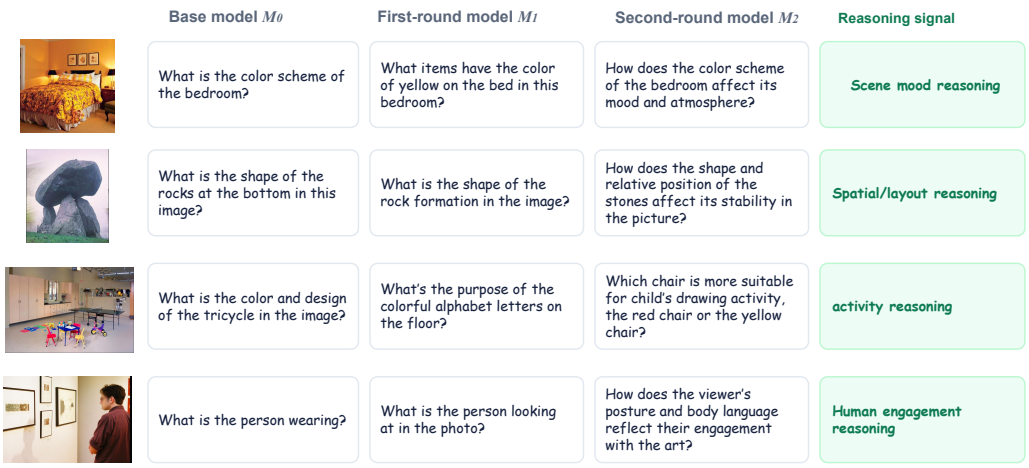

Figure 11: **Contextual reasoning in questions generated across rounds.** We compare the base (M_0), first-round (M_1), and second-round (M_2) questioners on the same images. Second-round questions connect visible cues to scene mood, physical stability, activities, and human engagement, illustrating a shift from attribute recognition toward contextual interpretation.