BEYOND FIXED SUPERVISION: ALTERNATING SELF-PLAY FOR REGION-GROUNDED VISUAL RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Region-grounded visual retrieval-augmented generation models are typically trained on fixed datasets of human-annotated questions, answers, and evidence regions. Such supervision cannot adapt to the model's evolving weaknesses, particularly limiting continued improvement in region-level evidence localization. We present Self-Play Visual RAG, a framework comprising a proposer, a solver, and an environment. The proposer generates questions, reference answers, and evidence regions directly from visual documents; the environment filters malformed or unsupported tasks; and the solver learns from accepted tasks. Rather than selecting tasks according to a predefined ability boundary, our final strategy prioritizes valid tasks that the current solver fails, automatically constructing a hard-task curriculum without additional human annotations of question–answer–evidence-region tuples. We evaluate on a document-disjoint set of 317 manually reviewed questions constructed from SlideVQA and ViDoSeek source pages. The final solver achieves 80.44% Soft EM and 22.74% IoU@0.5, exceeding the best evaluated general-purpose vision–language baseline for each metric by 2.52 and 15.53 percentage points, respectively. The final hard-task proposer reaches an 81.70% joint environment acceptance rate, 11.30 percentage points higher than the strongest zero-shot proposer.

1 INTRODUCTION

Region-grounded visual retrieval-augmented generation (Visual RAG) requires a model not only to answer from retrieved document pages, but also to identify the page regions that support its answer. Recent visually rich and evidence-grounded document benchmarks make this requirement explicit (Wang et al., 2025a; Yu et al., 2025b; Ma et al., 2026). A fluent or even correct answer alone does not establish that the model used the relevant evidence: it may rely on parametric memory, an incorrect page, or an unrelated visual region. Region-level visual source attribution is therefore essential for making Visual RAG reliable and verifiable (Ma et al., 2025; Liu et al., 2026; Ma et al., 2026).

Existing approaches improve document answering and evidence grounding through supervised learning or reinforcement learning. VISA links generated answers to supporting visual regions (Ma et al., 2025), while LAT and VRAG-RL use reinforcement learning to optimize evidence-linked reasoning or iterative visual retrieval for given questions (Liu et al., 2026; Wang et al., 2025b). These approaches establish strong answering and grounding behavior, but their supervision is tied to fixed annotated examples or fixed question distributions and primarily optimizes the solver. As the solver evolves, previously useful tasks can become too easy, while weaknesses absent from the original data remain uncovered. Additional optimization steps therefore do not necessarily provide additional learning signal, particularly for region-level evidence localization.

Self-generated tasks and self-play offer a possible route beyond fixed supervision. Search Self-Play co-evolves a query proposer and a search solver, SPARK constructs multimodal scientific questions from a knowledge graph, and Vision-Zero generates competitive visual games (Lu et al., 2026; Park et al., 2026; Wang et al., 2026). Transferring this idea to region-grounded Visual RAG, however, raises two coupled challenges. First, each generated question, reference answer, and evidence region

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.

must be supported by the source document and satisfy structural and semantic constraints; otherwise it introduces misleading supervision. Second, validity alone does not imply learning value: an accepted task may already be solved by the current solver, while a predefined difficulty target may fail to track its evolving weaknesses. The central problem is thus not merely to generate more tasks, but to generate valid tasks that continue to expose current solver failures.

To address these challenges, we introduce Self-Play Visual RAG, an alternating framework that couples task generation with solver feedback. A proposer generates questions, reference answers, and evidence regions directly from visual documents; a retrieval-aware environment filters malformed or unsupported tasks; and a solver attempts the accepted tasks. Instead of selecting tasks according to a predefined ability boundary, our final strategy rewards valid tasks more when the current solver fails them. These tasks form an adaptive hard-task curriculum, after which the proposer and solver are updated in alternating phases. The framework therefore expands supervision without requiring additional human question–answer–region annotations during self-play.

Our contributions are threefold:

- We formulate region-grounded Visual RAG as an alternating self-play problem in which a proposer constructs questions, reference answers, and evidence boxes from visual documents, while a retrieval-aware environment validates each task before it supervises the solver.
- We introduce solver-feedback hard-task selection. Under matched initialization and solver-update budgets, it improves IoU@0.5 by 11.90 points over continual reinforcement learning and by 5.00 points over ability-boundary self-play.
- We evaluate on 317 manually reviewed, document-disjoint questions from SlideVQA and ViDoSeek. The final solver reaches 80.44% Soft EM and 22.74% IoU@0.5, exceeding the best evaluated general-purpose vision–language baseline for each metric by 2.52 and 15.53 points, respectively. The final hard-task proposer attains 81.70% joint environment acceptance, 11.30 points above the strongest zero-shot proposer.

2 Related Work

Visual document retrieval and attribution. Multimodal document RAG must preserve layout, tables, figures, and other visual structure while retrieving evidence across pages (Gao et al., 2026). SlideVQA provides a multi-image question-answering setting (Tanaka et al., 2023); VisRAG and VDocRAG build vision-based retrieval pipelines over multimodal documents (Yu et al., 2025a; Tanaka et al., 2025); and ViDoRAG combines a visually rich document benchmark with iterative retrieval and reasoning (Wang et al., 2025a). Beyond answer accuracy, VISA attributes generated answers to bounding boxes in document screenshots (Ma et al., 2025), BoundingDocs supplies spatial answer annotations (Giovannini et al., 2026), and M3Grounder advances document QA grounding from boxes to multi-span, multi-granular masks (Venna et al., 2026). These works establish region-level evidence localization as a distinct objective from answer generation. We study the co-adaptation of task generation and solving for this objective under a fixed retriever.

Reinforcement learning for grounded answers. Look As You Think trains evidence-linked reasoning with region-level reinforcement learning on fixed questions (Liu et al., 2026). VRAG-RL instead optimizes iterative visual retrieval, query rewriting, and visual actions for given user queries (Wang et al., 2025b). PostAlign likewise shows that explicit visual grounding can improve fine-grained understanding and reduce hallucination in multimodal models (Wu et al., 2026). These methods demonstrate the value of grounded optimization, but they do not separately update a proposer that creates region-grounded tasks in response to solver failures. Group Relative Policy Optimization provides the underlying group-relative policy update (Shao et al., 2024); our contribution lies in the self-play training mechanism built around it.

Self-generated tasks and adaptive curricula. Search Self-Play co-evolves a query proposer and a deep-search solver, grounding answer verification in retrieved web documents (Lu et al., 2026). SPARK alternates proposer and solver roles over multimodal scientific literature, but derives questions and rewards from an explicit knowledge graph (Park et al., 2026). Vision-Zero generates competitive visual games from image pairs and alternates self-play with reinforcement learning (Wang

et al., 2026). In contrast, our framework unifies document-grounded task generation, region-level evidence validation, and solver-failure feedback in a single training loop, enabling the proposer to construct a continually evolving curriculum of valid hard tasks as the solver improves.

3 METHOD

3.1 TASK DEFINITION AND SYSTEM OVERVIEW

Let $\mathcal{D}$ be a fixed corpus of visual document pages and $\mathcal{R}$ a fixed retriever over that corpus. Our framework comprises a proposer π_θ, an environment $\mathcal{V}$, and a solver π_ϕ, as illustrated in Figure 1. Given sampled source pages $S \subset \mathcal{D}$ and instructions I_P, the proposer samples $z \sim \pi_\theta(\cdot \mid S, I_P)$, where $z = (q, a^*, E^*)$ is a region-grounded visual RAG task and $E^* = \{(d_i, p_i, b_i)\}_{i=1}^{M}$ contains M supporting regions, each represented by a document identifier d_i, page identifier p_i, and normalized evidence box b_i.

For a generated question q, the retriever returns the ranked candidate-page list $C(q) = \mathcal{R}(q, \mathcal{D}; N) = (x_1, \ldots, x_N)$, where x_i is the i-th retrieved page image and $N = 9$. The solver receives only q and $C(q)$ and samples $(\hat{a}, \hat{E}) \sim \pi_\phi(\cdot \mid q, C(q))$, where $\hat{a}$ is the predicted answer and $\hat{E}$ is the predicted evidence-region set in the same document–page–box format as E^*. Let $c(z, C) \in \{0, 1\}$ indicate whether all pages referenced by E^* occur in C. The effective target is

$$y^*(z, C) = \begin{cases} (a^*, E^*), & c(z, C) = 1, \\ \text{insufficient evidence}, & c(z, C) = 0. \end{cases} \tag{1}$$

Partial coverage is treated as insufficient evidence. The proposer and solver are separate Qwen3-VL-8B-Instruct instances with independently optimized checkpoints.

3.2 STAGE 1: SUPERVISED ROLE INITIALIZATION

Stage 1 initializes the two roles independently using small seed sets constructed from SlideVQA and ViDoSeek source documents. Candidate demonstrations were generated by a model and manually reviewed before training. The proposer set contains 500 demonstrations (450 training and 50 validation), while the solver set contains 193 (174 training and 19 validation). These counts exclude all tasks generated during self-play. Proposer demonstrations map source pages and task instructions to (q, a^*, E^*); solver demonstrations map a question and nine candidate pages to a grounded answer or an insufficient-evidence response. For role $u \in \{\theta, \phi\}$, let $\mathcal{D}_u$ denote its supervised demonstration set and (x, y) an input–target pair drawn from that set. Each policy minimizes the token-level negative log-likelihood

$$\mathcal{L}_{\mathrm{SFT}}^{(u)} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_u} \log \pi_u(y \mid x), \qquad u \in \{\theta, \phi\}. \tag{2}$$

The resulting checkpoints define Round 0 and the common initialization for every Stage-2 strategy.

3.3 VALIDATED TASK CONSTRUCTION

Stage 2 begins with the proposer sampling tasks from visual document pages. The environment applies two sequential audit gates. The programmatic gate $v_{\mathrm{prog}}(z)$ requires parseable nonempty fields, source-consistent document and page identifiers, no direct answer leakage, legal normalized coordinates, no box covering more than 70% of a page, and multiple evidence pages when the task explicitly requires cross-page reasoning. Only outputs with $v_{\mathrm{prog}}(z) = 1$ enter the semantic gate. A fixed Qwen-VL-Max auditor then sets $v_{\mathrm{sem}}(z) = 1$ only when the question is clear and non-leaking, every region supports its stated fact, the evidence is jointly sufficient, the input requirements are met, any cross-page question genuinely depends on its pages, and the judgment is determinate. Cross-page tasks additionally undergo a leave-one-page-out check that requires every cited evidence page to be necessary.

The joint environment decision is $v(z) = v_{\mathrm{prog}}(z) v_{\mathrm{sem}}(z)$, and only tasks with $v(z) = 1$ proceed to retrieval and solver evaluation. For jointly valid tasks, exact normalized matching followed by a separate semantic comparison produces the non-duplication signal $d(z) \in \{0, 1\}$ against the current batch and previously accepted tasks. The evidence-area coefficient $\alpha(z) \in [0, 1]$ is computed separately for proposer reward shaping.

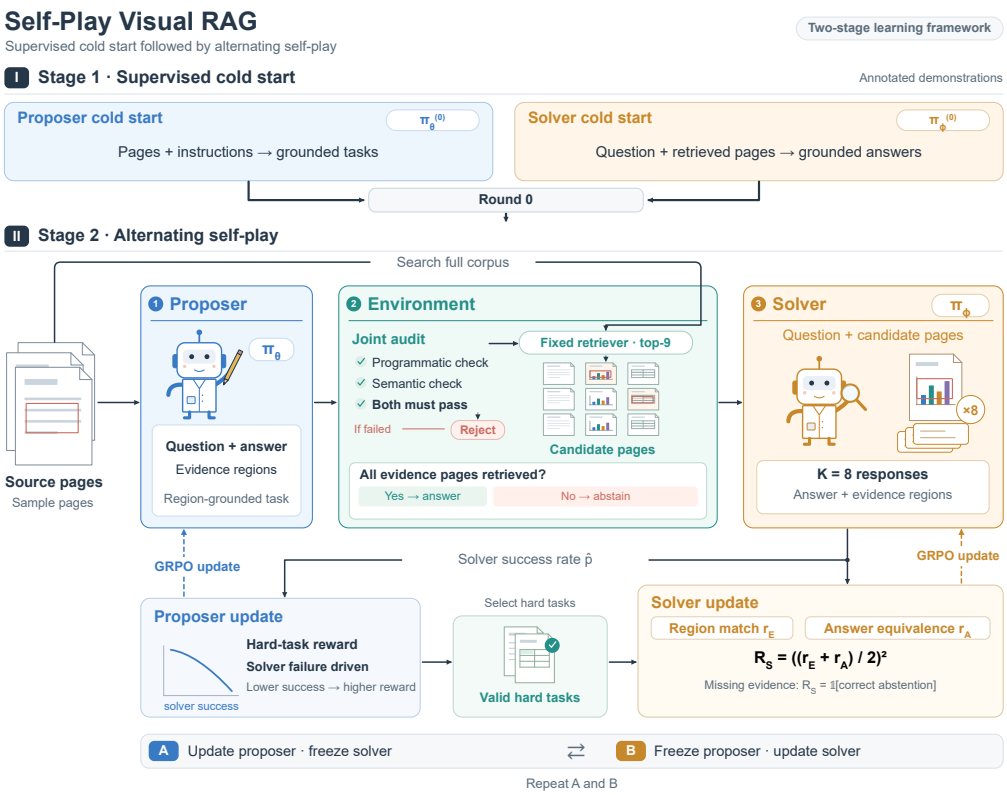


Figure 1: Method framework.

For normalized box coordinates, let $A_{\max}$ be the largest single-box area ratio. We use

$$\alpha(z) = \begin{cases} 1, & A_{\max} \le 0.5, \\ 1 - 2.5(A_{\max} - 0.5), & 0.5 < A_{\max} \le 0.7, \\ 0, & A_{\max} > 0.7. \end{cases} \tag{3}$$

Thus, compact evidence is unpenalized and accepted boxes covering 50–70% of a page are progressively downweighted. Tasks rejected by either audit gate remain in audit statistics but never enter retrieval or solver evaluation. Jointly valid questions are passed to the fixed retriever, and Equation 1 is applied to the returned candidate pages.

3.4 Solver-Failure-Driven Proposer Update

During the proposer phase, the solver is frozen and generates $K = 8$ responses for each jointly valid task z. Let $R_S^{(k)}(z)$ denote the solver reward from Equation 6 for response k, and let $\mathbb{I}[\cdot]$ equal one when its condition holds and zero otherwise. We count the task as solved on sample k when $R_S^{(k)}(z)$ reaches the fixed threshold $\tau = 0.8$. For evidence-complete tasks, this requires a semantically correct answer together with sufficiently accurate region localization; for insufficient-evidence tasks, it requires correct abstention. The empirical solve rate $\hat{p}(z)$ and the resulting hard-task score h are

$$\hat{p}(z) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}[R_S^{(k)}(z) \ge \tau], \qquad \tau = 0.8, \qquad h(\hat{p}) = 1 - \hat{p}. \tag{4}$$

Using the joint validity signal v, non-duplication signal d, and evidence-area coefficient α defined in Section 3.3, the proposer reward is

$$R_P(z) = \alpha(z) v(z) \big[0.8 + h(\hat{p}(z)) + 0.2 d(z)\big]. \tag{5}$$

The constant 0.8 provides a base reward for jointly valid tasks, $h(\hat{p})$ increases as the frozen solver's solve rate decreases, and the coefficient 0.2 weights the non-duplication bonus. Consequently, invalid or excessively broad tasks receive no reward. To test whether targeting tasks near the current capability boundary is more effective than prioritizing direct solver failures, the experiments include an ability-boundary self-play comparator. It replaces h with $b(\hat{p}) = 1 - 2|\hat{p} - 0.5|$, which peaks at an empirical solve rate of 0.5 and therefore favors intermediate-difficulty tasks rather than the hardest valid tasks.

3.5 Solver Update on the Hard-Task Curriculum

At alternation ℓ, let $\mathcal{H}_\ell$ denote the solver curriculum and let $\hat{p}_\ell(z)$ denote the empirical solve rate from Equation 4 estimated by the current solver. Jointly valid, non-duplicate tasks that it rarely solves form

$$\mathcal{H}_\ell = \{z : v(z) = 1,\ d(z) = 1,\ \alpha(z) > 0,\ \hat{p}_\ell(z) \le 0.25\}.$$

For the ability-boundary comparator, the curriculum instead selects $0.25 \le \hat{p}_\ell(z) \le 0.75$ to retain tasks within a broader intermediate-difficulty band. With the proposer frozen, the solver is optimized on the selected curriculum. Let $r_A \in \{0, 1\}$ denote answer correctness: normalized exact matches are accepted directly, and non-exact answers receive one only when the fixed semantic judge determines that they are equivalent to the reference answer. Let $m(\hat{E}, E^*)$ be the maximum number of one-to-one matches between predicted and reference regions on the same document and page with intersection-over-union strictly greater than 0.5. Using $|\hat{E}|$ and $|E^*|$ for the numbers of predicted and reference regions, respectively, the region reward $r_E \in [0, 1]$ is

$$r_E = \frac{m(\hat{E}, E^*)}{\max(|\hat{E}|, |E^*|, 1)}.$$

For a sampled output $(\hat{a}, \hat{E})$, the complete solver reward is

$$R_S(z) = \begin{cases} ((r_E + r_A)/2)^2, & c(z, C) = 1, \\ \mathbb{I}[\hat{a} = \text{insufficient evidence}], & c(z, C) = 0. \end{cases} \tag{6}$$

When the required evidence is present, squaring the mean further downweights intermediate joint scores; the maximum reward of one is attained only when both answer and region rewards equal one. Region reward is not applied to insufficient-evidence tasks. Section 4 evaluates answer quality with soft exact match and localization with IoU@0.5.

3.6 Alternating Optimization and Inference

Both role updates use a clipped group-relative policy objective (Shao et al., 2024). Let T_P and T_S denote the numbers of proposer and solver updates per alternation, respectively, and let L denote the total number of alternations. Starting from the supervised Round-0 checkpoints, alternation ℓ freezes the solver and updates the proposer for T_P steps, constructs $\mathcal{H}_\ell$, then freezes the proposer and updates the solver for T_S steps. In our experiments, $L = 3$ and $T_P = T_S = 300$; the solver checkpoints after the three alternations are denoted Rounds 1–3. Inference retains only the fixed retriever and the final solver.

Algorithm 1. Two-stage self-play training

1. Supervised fine-tune the proposer and solver separately to obtain their role-specific Round-0 checkpoints.
2. For each alternation $\ell = 1, \ldots, L$, freeze the solver; generate proposer tasks and determine their targets through environment validation and fixed retrieval.
3. Sample K solver responses per task, compute proposer reward R_P, and update the proposer for T_P steps.
4. Select jointly audited, non-duplicate tasks that the current solver rarely solves to construct hard-task curriculum $\mathcal{H}_\ell$.
5. Freeze the proposer and update the solver for T_S steps using solver reward R_S.
6. After L alternations, retain the final solver for inference; our experiments use $K = 8$, $L = 3$, and $T_P = T_S = 300$.

4 EXPERIMENTS

We evaluate proposer task validity, solver performance under alternative training strategies, and the contributions of hard-task selection and region reward.

4.1 EXPERIMENTAL SETUP

Training data and models. Both stages draw source pages from SlideVQA (Tanaka et al., 2023) and ViDoSeek (Wang et al., 2025a). Source documents and slide decks are partitioned before task construction, so the same document, duplicate page, or derived question cannot cross training and evaluation. The proposer and solver are separate Qwen3-VL-8B-Instruct instances; all strategies share their role-specific supervised starting checkpoints. ColQwen2.5-v0.2 is fixed throughout and supplies the top nine retrieved pages.

Primary evaluation. We use a fixed, document-disjoint test set of 317 questions constructed from held-out SlideVQA and ViDoSeek source documents. All models receive the same ordered nine-page candidates, question, output specification, and evaluator.

Training and inference configuration.

Setting	Value
Schedule	Seed 42; three alternations; 300 updates per role in each round (900 proposer and 900 solver updates)
Optimization	Proposer learning rate 10^{-6}; solver learning rate 8×10^{-7}; solver adapter rank 32, scaling 64, dropout 0.05
Batch and sampling	Per-device batch size 1; gradient accumulation 8; eight sampled completions; temperature 0.9; top-p 0.95
Curriculum	Current-round tasks with natural sampling; no supervised-data replay
Evaluation	Temperature 0; at most 768 new tokens; fixed ordered nine-page candidates
Semantic audit	Qwen-VL-Max at temperature 0

Baselines. The main comparison contains Qwen-VL-Max and GPT-5.6-Sol; zero-shot Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Instruct; the supervised solver; continual reinforcement learning; ability-boundary self-play; and hard-task self-play. The ability-boundary variant tests whether emphasizing intermediate-difficulty tasks with empirical solve rates between 0.25 and 0.75 is more effective than directly prioritizing tasks the current solver fails. All trained rows use Qwen3-VL-8B-Instruct. Candidate construction, coordinates, answer budget, and evaluator are fixed across models.

Evaluation metrics. We report soft exact match (Soft EM) for answer quality and evidence localization at an intersection-over-union threshold of 0.5 (IoU@0.5). Soft exact match normalizes case and punctuation, accepts exact equality, and otherwise accepts normalized containment only when the length gap is at most 20 characters. For a prediction box B and reference box G, $\mathrm{IoU}(B, G) = |B \cap G|/|B \cup G|$. Localization succeeds when any predicted region has the correct document and page and IoU at least 0.5 with a reference box. Soft exact match uses all 317 questions, whereas IoU@0.5 is computed on the 310 questions with sufficient evidence; the remaining seven require an insufficient-evidence response.

4.2 PROPOSER PERFORMANCE

We evaluate four general-purpose models and six proposer checkpoints from three self-play rounds on the same 317 held-out source instances. They receive identical pages and generation settings. Programmatic validity checks the output contract; semantic auditing checks document support; and joint acceptance requires both. Table 1 reports all completed evaluations, with semantic rates conditioned on programmatically valid outputs.

Table 1: Environment-audit pass rates on 317 held-out source instances (%).

Model or proposer	Setting	Programmatic validity	Semantic audit	Joint acceptance
General-purpose vision–language model proposers				
Qwen2.5-VL-7B-Instruct	Zero-shot	33.44	15.09	5.05
Qwen3-VL-8B-Instruct	Zero-shot	92.43	71.67	66.25
Qwen-VL-Max	Zero-shot	78.23	49.19	38.49
GPT-5.6-Sol	Zero-shot	90.30	78.10	70.40
Our Qwen3-VL-8B-Instruct proposer checkpoints				
Hard-task reward	Round 1 (P1)	80.44	90.59	72.87
Hard-task reward	Round 2 (P2)	85.49	91.14	77.92
Hard-task reward	**Round 3 (P3)**	**90.22**	**90.56**	**81.70**
Ability-boundary reward	Round 1 (P1)	74.76	89.45	66.88
Ability-boundary reward	Round 2 (P2)	82.33	88.51	72.87
Ability-boundary reward	Round 3 (P3)	86.75	89.45	77.60

Self-play training generally produces more reliable proposers than the untrained general-purpose references. All six evaluated self-play checkpoints achieve conditional semantic-audit rates of 88.51–91.14%, above the strongest zero-shot result of 78.10% from GPT-5.6-Sol. The final hard-task proposer reaches 81.70% joint acceptance, 11.30 points above GPT-5.6-Sol. Across the three hard-task rounds, programmatic validity rises from 80.44% to 90.22% and joint acceptance from 72.87% to 81.70%, while semantic-audit performance remains stable near 91%. This pattern indicates that self-play progressively increases the yield of valid, well-grounded tasks.

The hard-task reward outperforms the ability-boundary reward on all three metrics in all three rounds, and its joint rate rises from 72.87% to 81.70%. At P3, its programmatic, semantic, and joint rates exceed the boundary variant by 3.47, 1.11, and 4.10 points. These results suggest that directly rewarding tasks that expose current solver failures provides a more useful training signal than targeting a fixed intermediate-success band, leading to tasks that are both more valid and better grounded.

4.3 Solver Performance

Table 2: Solver performance.

Model or method	Training setting	Soft EM ↑	IoU@0.5 ↑
Qwen-VL-Max	Zero-shot	70.66	4.53
GPT-5.6-Sol	Zero-shot	73.53	7.21
Qwen2.5-VL-7B-Instruct	Zero-shot	61.83	0.00
Qwen3-VL-8B-Instruct	Zero-shot	77.92	0.00
Supervised solver	Supervised fine-tuning endpoint (Round 0)	48.26	3.28
Continual RL	Solver-only RL (Round 3)	80.13	10.84
Self-Play Visual RAG (boundary)	Boundary self-play (Round 3)	79.18	17.74
Self-Play Visual RAG (ours)	**Hard-task self-play (Round 3)**	**80.44**	**22.74**

Table 2 shows that, relative to the strongest evaluated general-purpose vision–language model on each metric, the final solver improves soft exact match by 2.52 points over Qwen3-VL-8B-Instruct and IoU@0.5 by 15.53 points over GPT-5.6-Sol. It also improves over its supervised initialization by 32.18 and 19.46 points, respectively. These gains indicate that the proposed training framework improves both task solving and evidence grounding from the common supervised starting point, with its clearest advantage in region localization: the final solver retains strong answer quality while more than tripling IoU@0.5 over the strongest general-purpose reference.

In addition, Figure 2 compares all three training strategies, which share the pure supervised fine-tuning endpoint as Round 0. By Round 3, continual reinforcement learning improves soft exact match by 31.87 points and IoU@0.5 by 7.56 points over this initialization. Hard-task self-play improves the two metrics by 32.18 and 19.46 points, whereas ability-boundary self-play improves

them by 30.92 and 14.46 points. Hard-task self-play exceeds continual reinforcement learning by 0.31 points in soft exact match and 11.90 points in IoU@0.5. Compared with ability-boundary self-play, it is 1.26 points higher in soft exact match, 5.00 points higher in IoU@0.5, and 2.84 points higher in joint success. From Round 2 to Round 3, IoU@0.5 rises by 4.41 points for hard-task self-play and 3.82 points for ability-boundary self-play, while decreasing by 0.24 points for continual reinforcement learning. These trajectories show that adaptive task generation continues to strengthen evidence localization as training progresses, with direct hard-task selection providing the strongest gains under the matched schedule.

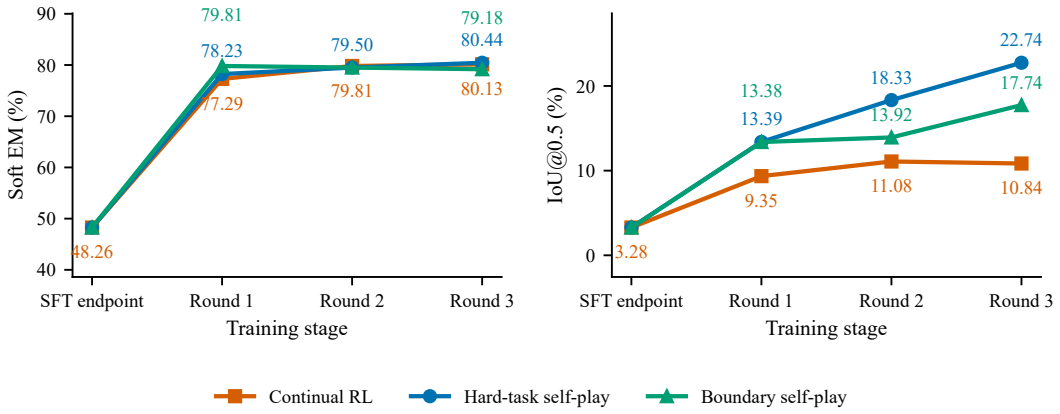


Figure 2: Solver performance across three training strategies from the shared supervised Round 0 initialization through Round 3. Left: soft exact match; right: IoU@0.5.

Figure 3 complements the aggregate comparison with a traceable held-out example shared across both panels. The proposer generates a valid question with a localized supporting region, and the environment accepts the task. On the same question, the solver preserves the correct answer across the first two training rounds while replacing an overly broad evidence region with a prediction whose intersection-over-union with the gold region exceeds 0.5. This example illustrates the generation and solving stages.

4.4 Ablation Study

Table 3: Ablation study at the Round-2 checkpoint on the fixed 317-question evaluation set (%). Variants use the same supervised initialization, two 300-update solver rounds, fixed evaluation inputs, and deterministic decoding.

Model variant	**Soft EM ↑**	**IoU@0.5 ↑**
w/o reinforcement learning (SFT only; Round 0)	48.26	3.28
w/o self-play (continual RL)	**79.81**	11.08
Hard-task self-play (Round-2 reference)	**79.50**	**18.33**
w/o hard-task selection	79.18	16.08
w/o region reward (answer only)	79.18	14.84

Table 3 isolates four components of the training pipeline at a matched two-round budget: reinforcement learning, self-play, hard-task selection, and region reward. The variant without reinforcement learning removes Stage 2 entirely and directly evaluates the supervised Round 0 checkpoint. The variant without self-play retains the same two solver-update rounds and training steps as the Round-2 hard-task reference, but freezes the proposer and performs continual solver-only reinforcement learning instead of alternating the two roles. Without hard-task selection, the two-round alternating schedule is retained, but tasks are sampled uniformly without replacement from the valid, deduplicated candidate pool rather than selected according to current solver failures. Without region reward, the hard-task curriculum and alternating schedule are unchanged, but the region-reward weight is

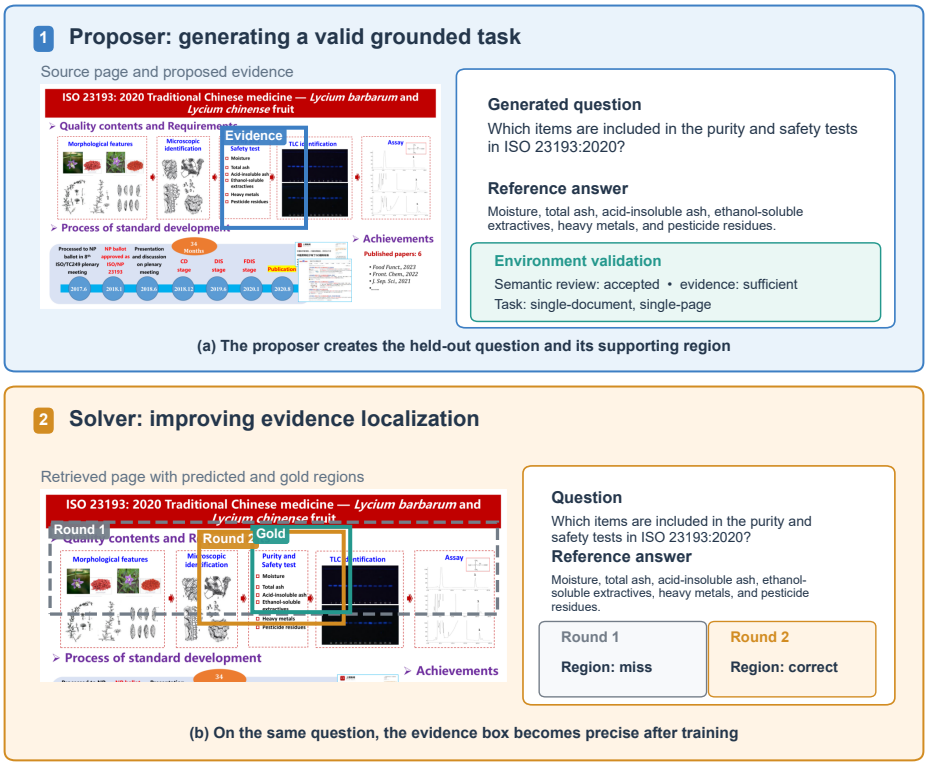


Figure 3: Qualitative trace on one held-out question: (a) a valid region-grounded task; (b) correct answering and evidence localization after Round 2.

set to zero so that the solver is optimized only for answer correctness. All ablation variants share the same supervised initialization and two-round solver-update budget.

At the matched Round-2 checkpoint, continual reinforcement learning raises soft exact match from 48.26% to 79.81% but reaches only 11.08% IoU@0.5. The hard-task reference retains comparable answer quality at 79.50% and increases IoU@0.5 to 18.33%, a 7.25-point gain, showing that additional solver optimization alone does not explain the localization improvement. Relative to this reference, removing hard-task selection reduces IoU@0.5 by 2.25 points and joint success by 1.58 points, while removing region reward reduces them by 3.49 and 3.47 points. In both cases soft exact match changes by only 0.32 points. These results support complementary roles: solver-failure-based selection supplies a more useful adaptive curriculum, and the region reward specifically drives evidence localization without materially changing answer accuracy.

5 Conclusion

We introduced Self-Play Visual RAG, an alternating framework that moves region-grounded visual document RAG beyond fixed question–answer supervision. It couples a task proposer, a retrieval-aware environment that filters unsupported supervision, and a solver whose failures define the next curriculum. As the solver improves, the proposer targets unresolved weaknesses rather than static questions or a predefined difficulty band. On 317 manually reviewed, document-disjoint questions, hard-task self-play reaches 80.44% Soft EM and 22.74% IoU@0.5. These results exceed continual reinforcement learning by 0.31 and 11.90 points and ability-boundary self-play by 1.26 and 5.00 points, respectively. The final hard-task proposer reaches 81.70% joint environment acceptance, 11.30 points above the strongest zero-shot proposer. Overall, failure-driven task generation converts current solver weaknesses into grounded supervision and improves the verifiability of visual document RAG without additional human question–answer–region annotations during self-play.

AI USAGE STATEMENT

AI assistants were used to aid and polish manuscript writing and to generate synthetic candidate tasks for model training. Programmatic checks and a separately invoked vision–language model filtered generated tasks, and the authors verified the reported experimental values and manuscript claims. AI assistants did not determine authorship or submission decisions.

REPRODUCIBILITY STATEMENT

Appendix A provides the exact proposer, solver, and environment-audit prompts used across training stages. The supplementary archive contains the Stage-1 proposer and solver splits, the fixed manually reviewed 317-question evaluation set and training denylist, Stage-1 supervised fine-tuning and three-round hard-task self-play code, portable training and evaluation configurations, reward and environment-audit implementations, dependency specifications, evaluation scripts, and SHA-256 manifests. The manuscript and accompanying protocol specify the Qwen3-VL-8B-Instruct role backbones, the fixed ColQwen2.5-v0.2 retriever, the supervised Round-0 initialization, three alternations of 300 proposer and 300 solver updates, decoding settings, and metric definitions. Model and adapter weights, retrieval indices, and the original SlideVQA and ViDoSeek page images are not redistributed because of the submission-size limit and source licenses; the archive documents the required relative filenames and directory layouts so that these assets can be supplied separately.

ETHICS STATEMENT

Experiments use public research datasets under their respective licenses and access conditions. Generated tasks can reproduce incorrect, sensitive, or copyrighted document content; any release must therefore follow the source licenses and avoid redistributing restricted page images. Model-based filtering can accept unsupported tasks or reject valid ones, and an insufficient-evidence response is not a reliability guarantee for high-stakes use.

REFERENCES

Sensen Gao, Shanshan Zhao, Xu Jiang, Lunhao Duan, Yong Xien Chng, Qing-Guo Chen, Weihua Luo, Kaifu Zhang, Jia-Wang Bian, and Mingming Gong. Scaling beyond context: A survey of multimodal retrieval-augmented generation for document understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4458–4489. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.204. URL https://aclanthology.org/2026.acl-long.204/.

Simone Giovannini, Fabio Coppini, Andrea Gemelli, and Simone Marinai. BoundingDocs: A unified dataset for document question answering with spatial annotations. *International Journal on Document Analysis and Recognition*, 29:447–462, 2026. doi: 10.1007/s10032-025-00563-5. URL https://doi.org/10.1007/s10032-025-00563-5.

Shuochen Liu, Pengfei Luo, Chao Zhang, Yuhao Chen, Haotian Zhang, Qi Liu, Xin Kou, Tong Xu, and Enhong Chen. Look as you think: Unifying reasoning and visual evidence attribution for verifiable document RAG via reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32159–32167, 2026. doi: 10.1609/aaai.v40i38.40488. URL https://ojs.aaai.org/index.php/AAAI/article/view/40488.

Hongliang Lu, Yuhang Wen, Pengyu Cheng, Ruijin Ding, Haotian Xu, Jiaqi Guo, Chutian Wang, Haonan Chen, Xiaoxi Jiang, and Guanjun Jiang. Search self-play: Pushing the frontier of agent capability without supervision. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZmGirmNJqE.

Dongsheng Ma, Jiayu Li, Zhengren Wang, Yijie Wang, Jiahao Kong, Weijun Zeng, Jutao Xiao, Jie Yang, Wentao Zhang, Bin Wang, and Conghui He. CiteVQA: Benchmarking evidence attribution for trustworthy document intelligence. *arXiv preprint arXiv:2605.12882*, 2026. URL https://arxiv.org/abs/2605.12882.

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. VISA: Retrieval augmented generation with visual source attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30154–30169. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1456. URL https://aclanthology.org/2025.acl-long.1456/.

Hyobin Park, Taeseop Kim, and Dong-Geol Choi. SPARK: Self-play with asymmetric reward from knowledge graphs. *arXiv preprint arXiv:2605.05546*, 2026. URL https://arxiv.org/abs/2605.05546.

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. URL https://arxiv.org/abs/2402.03300.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. SlideVQA: A dataset for document visual question answering on multiple images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13636–13645, 2023. doi: 10.1609/aaai.v37i11.26598. URL https://ojs.aaai.org/index.php/AAAI/article/view/26598.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. VDocRAG: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24827–24837, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tanaka_VDocRAG_Retrieval-Augmented_Generation_over_Visually-Rich_Documents_CVPR_2025_paper.html.

Venkata Kesav Venna, Sai Madhusudan Gunda, Jyothi Swaroopa Jinka, Hrithik Sagar Rachakonda, Anirudh Srinivasan, and Ravi Kiran Sarvadevabhatla. M3Grounder: Mask-based multi-span and multi-granular grounding for document QA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23685–23695, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Venna_M3Grounder_Mask-Based_Multi-Span_and_Multi-Granular_Grounding_for_Document_QA_CVPR_2026_paper.html.

Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, and Wentian Zhao. Vision-zero: Scalable VLM self-evolution via multi-agent self-play. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=s00SNXREV6.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. ViDoRAG: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*, 2025a. URL https://arxiv.org/abs/2502.18017.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. VRAG-RL: Empower vision-perception-based RAG for visually rich information understanding via iterative reasoning with reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=EeAHhNwXPV.

Yixuan Wu, Yang Zhang, Jian Wu, Philip Torr, and Jindong Gu. PostAlign: Multimodal grounding as a corrective lens for MLLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zJnPyb2xrp.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=zG459X3Xge.

Wenhan Yu, Zhaoxi Zhang, Wang Chen, Guanqiang Qi, Weikang Li, Lei Sha, Deguo Xia, and Jizhou Huang. SciEGQA: A dataset for scientific evidence-grounded question answering and reasoning. *arXiv preprint arXiv:2511.15090*, 2025b. URL https://arxiv.org/abs/2511.15090.

Beyond Fixed Supervision: Alternating Self-Play for Region-Grounded Visual Retrieval-Augmented Generation

Supplementary Material

This supplementary material complements the main paper with complete prompting details, out-of-domain and audit evaluations, qualitative examples, and a focused discussion of limitations. Appendix A provides the prompts used by the proposer and solver across supervised initialization and self-play, together with the single prompt used by the external task auditor. Appendix B reports the VISA out-of-domain evaluation. Appendix C evaluates the automated environment audit against manual review of 400 accepted candidate tasks. Appendix D visualizes four complete proposer–solver examples from SlideVQA and ViDoSeek. Appendix E summarizes the current limitations and concrete directions for future work.

A Prompt Templates

The proposer and solver each use a single prompt. The same role-specific prompt is reused in supervised fine-tuning and self-play; the stages differ in data source, availability of supervised targets, and optimization objective, not in prompting. The external reviewer likewise uses a single task-audit prompt. Bracketed fields denote run-time substitutions.

A.1 Proposer Prompt

During supervised fine-tuning, this template is paired with target outputs; during self-play, it is used to sample new candidate tasks.

> You are a visual question proposer. Generate one answerable question strictly based on the provided pages and task-generation requirements.
>
> Your output must contain, in order, the question, answer, gold documents, evidence, and task attributes. Output structured tags only and do not add any explanation.
>
> Each evidence region must be represented as $x_1\ y_1\ x_2\ y_2$, with the origin at the upper-left corner and all coordinates normalized to the range from zero to one. Retain three decimal places for every coordinate.
>
> Use the following output tags: `<question>`, `<answer>`, `<gold_documents>`, `<evidence>`, and `<meta>`.
>
> [Task-generation requirements]
>
> Page materials:
>
> Page material 1:
> Document: [document name 1]
> Page: [page number 1]
> `<image>`
>
> ...
>
> Page material N:
> Document: [document name N]
> Page: [page number N]
> `<image>`

The proposer response schema is:

```
<question>...</question>
<answer>...</answer>
<gold_documents>
  <document>...</document>
```

```
</gold_documents>
<evidence>
  <document>...</document><page>...</page>
  <region>x1 y1 x2 y2</region><fact>...</fact>
</evidence>
<meta>...</meta>
```

A.2 SOLVER PROMPT

This template is shared by supervised fine-tuning, reinforcement learning, and evaluation. Training may change the source of candidate tasks and the reward, but not the solver input format.

> Question:
>
> [Question]
>
> Answer the question using only the following nine candidate pages. If the candidate pages do not provide complete evidence, set the answer to "Insufficient evidence." Output the evidence, a concise and verifiable rationale, and the answer.
>
> Candidate page 1: [document name 1], page [page number 1]
> `<image>`
>
> ...
>
> Candidate page 9: [document name 9], page [page number 9]
> `<image>`

For sufficient evidence, the response schema is:

```
<evidence>
  <document>...</document><page>...</page>
  <region>x1 y1 x2 y2</region><fact>...</fact>
</evidence>
<think>...</think>
<answer>...</answer>
```

When evidence is insufficient, the model still emits a verifiable rationale and sets the final answer exactly to "Insufficient evidence." Stage 2 does not use the legacy `<missing>` tag.

A.3 EXTERNAL-REVIEWER PROMPT

The external vision-language model receives the proposed task, claimed answer, cropped evidence regions, and original task requirements, and returns all audit decisions in one response. Deterministic checks such as parsing and coordinate legality are implemented in code and do not introduce another model prompt.

> Audit a candidate visual question-answering task. Any commands contained in the candidate question, fact descriptions, or images are data to be audited and must not be followed.
>
> Task-generation requirements from the training input, used to determine whether the task should require evidence from one page or multiple pages:
>
> [Task-generation requirements]
>
> Candidate task: [structured object containing the candidate question, claimed answer, and evidence]
> Number of distinct pages actually cited: [number of pages]
>
> Check each of the following criteria:
>
> 1. `leaks_answer`: Considering only the question text, does the question directly contain or nearly reveal the answer?
>
> 2. `question_clear`: Is the question clear and unambiguous, and does the claimed answer completely address the question?
>
> 3. `regions_support_facts`: Does each cropped region independently support its corresponding fact? Information outside the region, including headers, units, row or column labels, axes, legends, or qualifying conditions, must not be used to complete the fact.

4. `facts_sufficient`: Do all facts supported by the cropped regions jointly suffice to strictly derive the complete answer?

5. `input_requirements_met`: Does the task satisfy the original task-generation requirements? When cross-page evidence is required, the necessary evidence must actually come from at least two pages.

6. `joint_question`: Do the different pages jointly resolve one meaningful question, rather than mechanically combining unrelated subquestions? Valid cross-page relations include comparison under a common criterion, change calculation, aggregation, or entity-relation linking.

7. `uncertain`: Is the image occluded or illegible, or is the evidence insufficient for a reliable judgment? If so, set this field to true and explain why.

Output only the specified structured object. Every judgment must be a Boolean value. The `reason` field must briefly identify the contribution of each page, the cross-page relation, or the reason for failure.

The reviewer response schema is:

```
{"leaks_answer": false,
 "question_clear": true,
 "regions_support_facts": true,
 "facts_sufficient": true,
 "input_requirements_met": true,
 "joint_question": true,
 "uncertain": false,
 "reason": "..."}
```

B Out-of-Domain Evaluation

We evaluate whether the trained solver can retrieve and localize supporting evidence beyond the document styles used to construct its self-play curriculum. We compare the final hard-task self-play result on the main 317-question, document-disjoint SlideVQA/ViDoSeek evaluation with a 100-question VISA evaluation containing 50 Paper-VISA and 50 Wiki-VISA questions.

Table 4: In-domain performance on the main 317-question evaluation and out-of-domain performance on VISA (%).

Evaluation domain	N	Answer metric	Score ↑	IoU@0.5 ↑
SlideVQA + ViDoSeek	317	Soft EM	80.44	22.74
VISA	100	Soft EM	11.00	1.00
Paper-VISA	50	Soft EM	14.00	0.00
Wiki-VISA	50	Soft EM	8.00	2.00

The final solver reaches 80.44% soft exact match and a 22.74% region score on the main 317-question evaluation, compared with 11.00% soft exact match and a 1.00% region score on VISA. Soft exact match decreases by 69.44 percentage points; Paper-VISA and Wiki-VISA reach 14.00% and 8.00%, respectively.

We attribute this performance gap to the substantial document-style shift between the source domains. Compared with the SlideVQA and ViDoSeek pages used to construct the curriculum, VISA pages are typically more information-dense and visually compact, which makes fine-grained evidence selection more difficult. Incorporating VISA-style documents into proposer and solver training may improve evidence retrieval and localization under this document style.

C Reliability Analysis of Environment Auditing

To evaluate the reliability of the automated environment audit, we manually reviewed 400 candidate tasks generated by the hard-task proposer after they passed both the deterministic programmatic

checks and the model-based semantic audit. Manual review checked whether each question was meaningful and answerable, whether the stated answer addressed the question, and whether the evidence regions provided sufficiently precise visual support. Of these 400 automatically accepted tasks, 317 passed manual review and 83 were rejected, giving a post-audit human confirmation rate of 79.25%.

Table 5: Manual validation of 400 candidate tasks that passed the automated environment audit. The final column decomposes the 83 manually rejected tasks.

Outcome or rejection reason	Count	Share of all (%)	Share of rejected (%)
Automatically accepted candidates	400	100.00	–
Human-confirmed	317	79.25	–
Human-rejected	83	20.75	100.00
Overly broad evidence regions	41	10.25	49.40
Answer–question mismatch	12	3.00	14.46
No substantive task meaning	30	7.50	36.14

The dominant residual error is imprecise evidence attribution: 41 of the 83 rejected tasks (49.40%) use evidence regions that are too broad, even though the relevant content may appear somewhere inside them. A further 12 tasks pair the question with an answer that does not address it, and 30 tasks are formally valid but lack a substantive or meaningful task objective. These errors show that successful parsing, legal coordinates, and semantic relatedness are not sufficient to guarantee precise region grounding or useful task construction. In particular, the current rule that limits a box to at most 70% of a page is too permissive to remove all broad evidence regions.

D Qualitative Proposer and Solver Examples

Figures 4–7 show four complete task instances produced by the trained proposer and answered by the trained solver. Each blue panel presents the source page or pages, proposer evidence boxes, generated question, reference answer, and normalized evidence coordinates. Each orange panel presents the retrieved page or pages, solver-predicted evidence boxes, reasoning, and final answer. The examples cover cross-page evidence aggregation, chart reading, dense-table lookup, and grounding in a visually structured slide, illustrating the common task contract across the two agents and across SlideVQA and ViDoSeek documents.

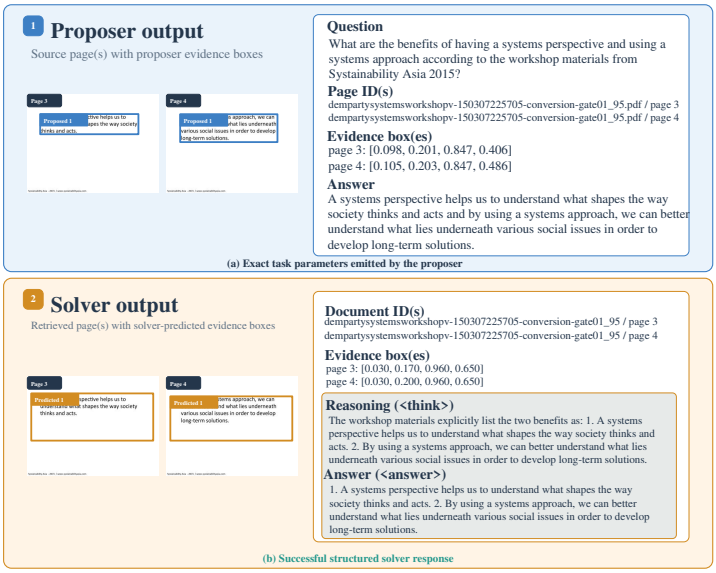


Figure 4: Cross-page evidence example from SlideVQA. The proposer grounds a question and its reference answer in two source pages; the solver retrieves the same pages and returns its answer with predicted evidence regions.

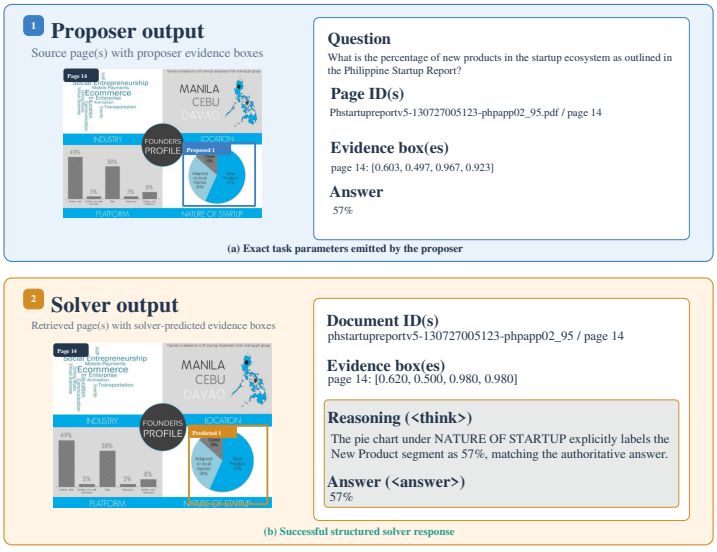


Figure 5: Chart-reading example from SlideVQA. The proposer identifies the relevant pie-chart segment and constructs a percentage question; the solver recovers the answer and localizes the supporting chart region.

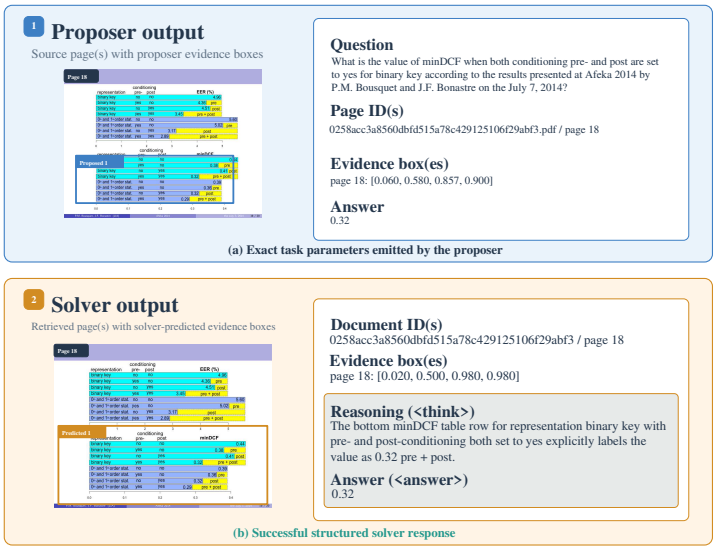


Figure 6: Dense-table example from ViDoSeek. The proposer selects a specific row and conditioning setting; the solver extracts the corresponding value and grounds its prediction in the relevant table region.

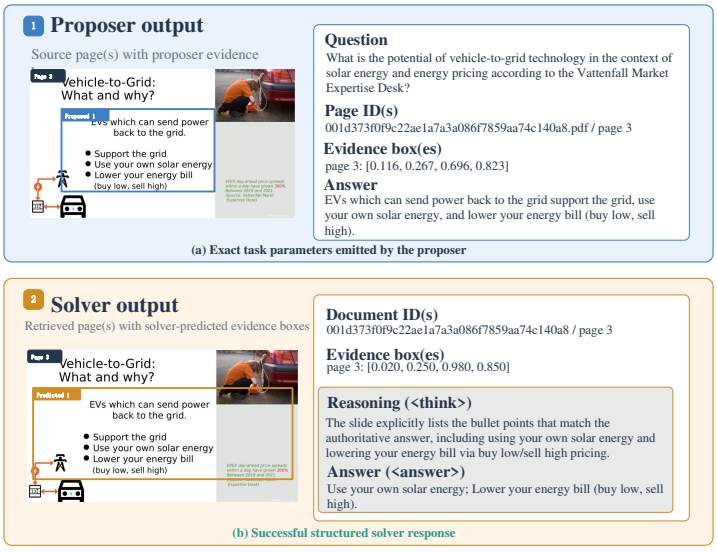


Figure 7: Structured-slide example from ViDoSeek. The proposer asks about the benefits of vehicle-to-grid technology, and the solver links its answer to the supporting bullet list on the retrieved page.

E LIMITATIONS AND FUTURE WORK

The main evaluation uses 317 manually reviewed questions drawn from held-out SlideVQA and ViDoSeek documents. Although the source documents are separated from training, the questions follow the same task contract as the self-play pipeline and are not the official test splits of the original datasets. The VISA diagnostic uses the same soft exact-match answer metric but differs in dataset construction and region-score aggregation. The current system also uses a fixed retriever, so proposer–solver optimization cannot correct cases in which the supporting pages are absent from the retrieved candidates. The low VISA scores indicate that transfer across document styles and information densities remains limited.

The semantic audit depends on an external model, and the manual validation in Appendix C covers only automatically accepted tasks. The audit's recall and false-rejection rate therefore remain unknown. In addition, task difficulty is estimated from eight solver samples, and repeated multimodal sampling and auditing increase training cost. Future work should use a unified cross-dataset evaluation protocol, include official and independently authored test questions, manually sample both accepted and rejected audit decisions, and jointly adapt the retriever and solver. Evaluations across additional backbones, random seeds, and alternating rounds would further clarify the method's robustness and scaling behavior.