VisPlay: Self-Evolving Vision-Language Models

Yicheng He[1*] Chengsong Huang[2*] Zongxia Li[3*] Jiaxin Huang[2] Yonghui Yang[4,5†]
[1]University of Illinois Urbana-Champaign [2]Washington University in St. Louis
[3]University of Maryland [4]National University of Singapore
[5]Intelligent Interconnected Systems Laboratory of Anhui Province, Hefei University of Technology
yh84@uiuc.edu chengsong@wustl.edu zli12321@umd.edu

Abstract

Reinforcement learning (RL) provides a principled framework for improving Vision-Language Models (VLMs) on complex reasoning tasks. However, existing RL approaches often depend on human-annotated labels or task-specific heuristics to define verifiable rewards—both costly and limited in scalability. We introduce VisPlay, a self-evolving RL framework that enables VLMs to autonomously improve their reasoning capabilities from massive unlabeled image data. Starting from a single base VLM, VisPlay assigns the model into two interacting roles: an ***Image-Conditioned Questioner*** *that formulates challenging yet answerable visual questions, and a* ***Multimodal Reasoner*** *that generates silver responses. These roles are jointly trained using Group Relative Policy Optimization (GRPO), which uses diversity and difficulty rewards to balance the difficulty of generated questions with the quality of silver answers. VisPlay scales efficiently across two model families. Trained on Qwen2.5-VL and MiMo-VL, VisPlay achieves consistent improvements in visual reasoning, compositional generalization, and hallucination reduction across eight benchmarks including MM-Vet and MMMU, and establishes a scalable path toward self-evolving multimodal intelligence. Our code is available at* `https://github.com/bruno686/VisPlay`.

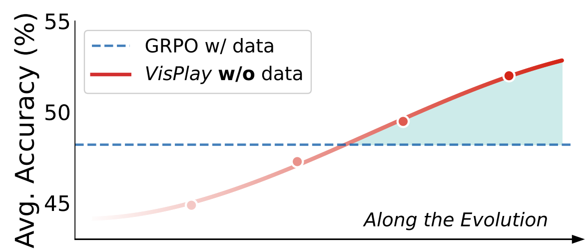


Figure 1. Illustration of the average accuracy improvement (averaged over seven datasets) through successive evolutions (Evo 1 to Evo 5) on Qwen2.5-VL-3B-Instruct, compared to a baseline trained on Vision-47K with GRPO, demonstrating the effectiveness of our *VisPlay* .

1. Introduction

Self-evolving mechanisms [6, 22, 29, 39] represent a promising frontier for advancing artificial intelligence. The training of state-of-the-art (SoTA) models has traditionally relied on large volumes of expert curated tasks and labels. However, the reliance on human annotation is not only costly, labor-intensive, and difficult to scale, but also presents a fundamental bottleneck to advancing intelligence toward capabilities that could surpass itself without human signal guidance [42]. Self-evolution offers a compelling alternative by equipping models with the capacity to independently generate, refine, and learn from their own experiences such as through self-play or synthetic data generation.

Motivated by these advantages, the research community has increasingly explored self-evolution, most notably in the context of Large Language Models (LLMs). A line of works have demonstrated how LLMs can autonomously enhance their complex reasoning and coding faculties, often by generating their own tasks or data [13, 24, 50]. However, the self-evolution paradigm remains largely underexplored for Vision-Language Models (VLMs) [2, 18, 32]. Unlike LLMs, which rely solely on text, developing self-evolving VLMs poses additional challenges due to their dependence on the visual modality. In a world where human annotation is costly and time-consuming, yet vast amounts of visual data are freely available online, self-evolving VLMs present a promising direction to continual improvement without human signals and directly from the abundant visual content on the internet [3, 30].

In this paper, we introduce *VisPlay*, a self-evolving RL framework that enables VLMs to autonomously improve their reasoning capabilities using **only raw, unanno-**

*Equal Contribution.
†Corresponding Author yh_yang@nus.edu.sg.

tated images. The framework utilizes a single base VLM that alternates between two roles: the *Image-Conditioned Questioner*, which generates diverse and challenging questions conditioned on an input image, and the *Multimodal Reasoner*, which produces silver responses based on both the image and the generated question. Both roles are jointly optimized using Group Relative Policy Optimization (GRPO) [31], where designed rewards encourage a balance between question difficulty and answer quality without requiring external supervision. The Image-Conditioned Questioner learns to generate challenging yet answerable questions grounded in visual inputs, while the Multimodal Reasoner learns to produce accurate, detailed, and grounded responses. This self-evolving framework enables the VLM to progressively improve its visual reasoning abilities through iterative co-improvement of the Questioner and the Reasoner as Figure 1.

We apply our self-evolving RL framework to train three state-of-the-art (SoTA) VLMs and observe consistent performance gains across diverse visual reasoning benchmarks. Our main contributions are:

- We propose *VisPlay*, a self-evolving RL framework for Vision-Language models.
- We apply *VisPlay* to three strong models—Qwen2.5-VL-3B, Qwen2.5-VL-7B [32], and MiMo-VL-7B [40]. We run extensive evaluations over three major domains—General Visual Understanding, Visual Mathematics, and Hallucination Detection. All models show consistent gains in accuracy after several iterations.
- We run extensive ablation studies to further validate the contribution of Image-Conditioned Questioner and the Multimodal Reasoner component to further show how *VisPlay* progressively strengthens multimodal reasoning across vision-language tasks.

2. Method

2.1. Preliminary

Reinforcement Learning with Verifiable Rewards (RLVR) [16] is a paradigm for training VLMs in domains where the correctness of model outputs can be verified. A rule-based verifier $v : X \to \{0, 1\}$ assigns a binary reward to each generation x_i:

$$r_i = v(x_i) = \begin{cases} 1, & \text{if } x_i \text{ satisfies a correctness rule,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Such verifiable rewards are effective in tasks like mathematical reasoning, multiple choice, and code generation, where correctness can be objectively evaluated. GRPO [31] provides a practical RL algorithm without a value function by using relative rewards among multiple samples from the same prompt. Given a prompt p, a policy $\pi_{\theta_{\text{old}}}$ produces G complete responses $\{x_1, \ldots, x_G\}$ with corresponding rewards $\{r_1, \ldots, r_G\}$. Rewards are normalized within the group to compute response-level advantages:

$$\hat{A}_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G) + \varepsilon_{\text{norm}}}, \tag{2}$$

where $\varepsilon_{\text{norm}}$ is a small constant for stability.

The policy is then optimized using a clipped surrogate objective, regularized by a KL term to constrain policy drift:

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G} \min\Bigg(& \frac{\pi_\theta(x_i)}{\pi_{\theta_{\text{old}}}(x_i)} \hat{A}_i, \\ & \text{clip}\left(\frac{\pi_\theta(x_i)}{\pi_{\theta_{\text{old}}}(x_i)}, 1-\epsilon,\ 1+\epsilon\right) \hat{A}_i \Bigg) \\ & + \beta\, \text{KL}(\pi_\theta \,\|\, \pi_{\theta_{\text{old}}}) \,. \end{aligned} \tag{3}$$

GRPO operationalizes RLVR to improve reasoning and generation quality in VLMs by rewarding responses with positive relative advantages while limiting policy deviation.

2.2. Pipeline Overview

We introduce *VisPlay*, a self-play reinforcement learning framework designed to evolve VLMs without human-annotated data. As illustrated in Figure 2, the framework operates as a closed-loop system involving two agents evolved from the same base model: an *Image-Conditioned Questioner* and a *Multimodal Reasoner*. The process begins with the Questioner taking an image as input to generate a visual query. Subsequently, the Reasoner receives both the image and the generated query to produce a response. Both the Questioner and the Reasoner are initialized from a shared pretrained backbone. The two agents co-evolve through iterative interactions: the Questioner is trained to generate more challenging questions, while the reasoner is trained to solve more and more challenging questions. The complete process is described in Algorithm 1.

2.3. Image-Conditioned Questioner Training

The Questioner is an autoregressive policy denoted by Q_θ. Conditioned on an input image I, it samples a group of G questions $\{x_i\}_{i=1}^{G} \sim Q_\theta(\cdot|I)$, which are evaluated to produce scalar rewards $\{r_i\}_{i=1}^{G}$. These rewards are used to compute group-normalized advantages and to update Q_θ with a GRPO objective. We next define reward components that constitute each r_i.

Pseudo-Label Generation. Since self-evolving VLMs learn without relying on labeled data, ground-truth answers for the Questioner's generated questions are unavailable.

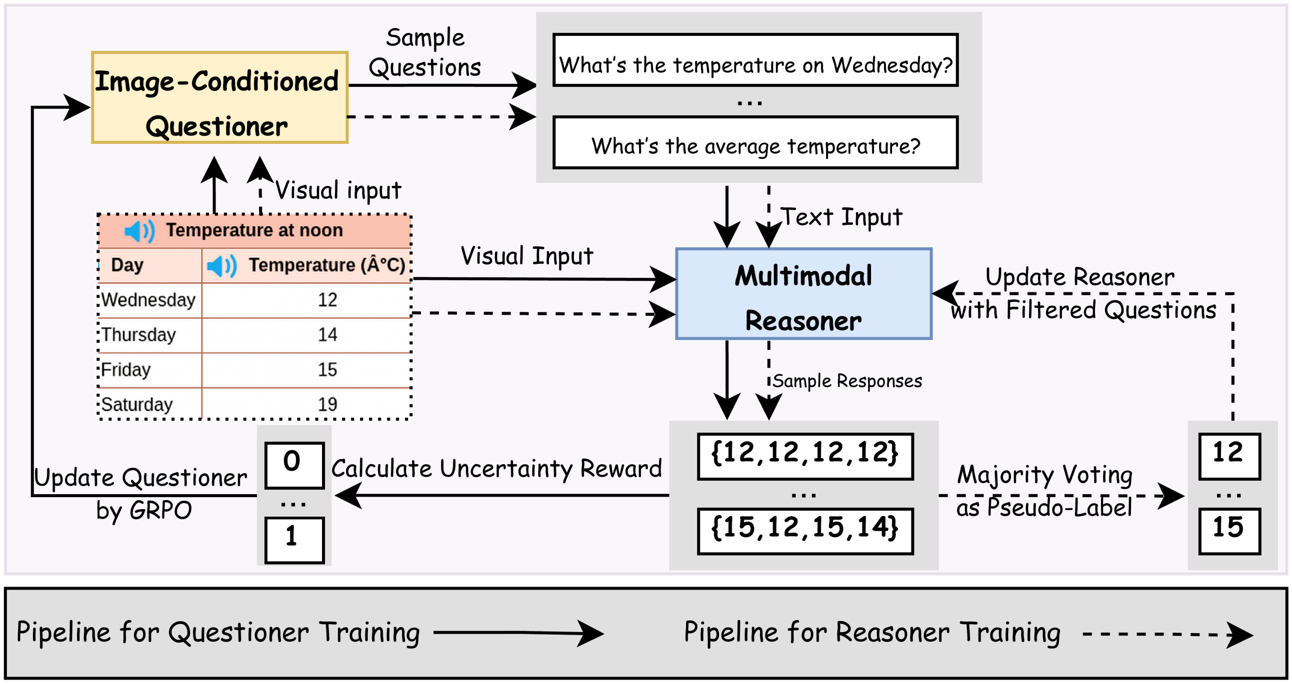


Figure 2. An illustration of our *VisPlay* framework, depicting the co-evolution of the Image-Conditioned Questioner and Multimodal Reasoner. Top: During the Questioner training stage, the Image-Conditioned Questioner is optimized via GRPO to produce challenge questions. The reward stems from the uncertainty of the frozen Multimodal Reasoner, computed by the consistency of its multiple generated answers. Bottom: In the Reasoner training stage, the Multimodal Reasoner is trained via GRPO on a curated set of challenging questions from the now-frozen Image-Conditioned Questioner, leveraging pseudo-labels from its own majority voting.

Therefore, we introduce a method to approximate the corresponding answers. Given an image I and the generated question x, we introduce a Reasoner S_ϕ that samples m responses $\{y_j\}_{j=1}^m$.[1] We define the empirical frequency of a candidate answer y as $\hat{p}(y|x,I) = \frac{1}{m}\sum_{j=1}^m \mathbb{1}\{y_j = y\}$, and derive the pseudo-label via majority voting: $\tilde{y} = \arg\max_y \hat{p}(y|x,I)$. We then define the **confidence score** for this pseudo-label as:

$$\mathrm{conf}(x,I) = \hat{p}(\tilde{y}|x,I). \tag{4}$$

Intuitively, $\mathrm{conf}(x,I)$ measures the Reasoner's certainty about the pseudo-label: high values indicate stable and consistent predictions, whereas values near 0.5 reflect strong uncertainty. We therefore treat the degree of uncertainty (i.e., how close $\mathrm{conf}(x,I)$ is to 0.5) as a proxy for the model-perceived difficulty of the generated question.

Uncertainty Reward. The confidence score quantifies the Reasoner's uncertainty, which we use as a proxy for the model-perceived difficulty of the generated question. To encourage questions that probe the Reasoner's limits, we compute the reward based on the confidence score $c = \mathrm{conf}(x,I)$. We define the uncertainty reward to penalize deviations from the point of maximum uncertainty:

$$r_{\mathrm{unc}}(x,I) = 1 - \big|2c - 1\big|. \tag{5}$$

This formulation yields a maximal reward of 1 when $c = 0.5$ and decreases linearly to 0 as the reasoner's response distribution becomes deterministic (i.e., $c \to 1$).

Diversity Regularization. To prevent the model from collapsing into generating repetitive questions for a given image I, we introduce a redundancy penalty within its generated group $\mathcal{X}_I$. We cluster these generated questions based on pairwise similarity (BLEU score) to identify duplicates. For a question x_i belonging to a cluster $C_k^{(I)} \subseteq \mathcal{X}_I$, the regularization term is:

$$r_{\mathrm{div}}(x_i,I) = \lambda \frac{|C_k^{(I)}|}{G}, \tag{6}$$

where $C_k^{(I)}$ denotes the cluster of similar questions for image I, and G is the total number of generated questions for that image.

Format constraint. We enforce a hard filter to ensure structural validity. Specifically, we require the generated

[1] The reasoner is evolved from the same base model as the Questioner. See Section 2.4 for details.

question to be strictly enclosed within <question> tags. Any output failing to meet this format requirement is assigned zero reward. We denote this validity indicator as:

$$\mathbb{1}_{\text{valid}}(x) = \begin{cases} 1, & \text{if } x \text{ is wrapped in <question> tags,} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Final Questioner Reward. For each generated question x_i conditioned on image I, we integrate the uncertainty signal and diversity regularization into a unified scalar objective:

$$r_i = \mathbb{1}_{\text{valid}}(x_i) \cdot \text{ReLU}\Big(r_{\text{unc}}(x_i, I) - r_{\text{div}}(x_i, I)\Big). \tag{8}$$

This composite reward incentivizes the Questioner to generate challenging yet non-redundant questions while strictly filtering out malformed outputs. The ReLU function stabilizes GRPO updates by preventing spurious negative values from skewing the reward normalization across the group.

2.4. Multimodal Reasoner Training

The training of the Multimodal Reasoner S_ϕ builds upon the advancements of the Image-Conditioned Questioner. In each iteration, the Image-Conditioned Questioner functions produce challenging samples that serve as training targets. The Multimodal Reasoner then learns from these automatically curated samples, improving its visual reasoning ability without any external supervision.

Curated Dataset Construction. Following the update of the Image-Conditioned Questioner, we generate a diverse pool of N candidate questions $\{x_i\}_{i=1}^N$ per image by sampling $x_i \sim Q_\theta(\cdot \mid I)$. For each x_i, we obtain m response samples from the current Multimodal Reasoner and compute the pseudo-label $\tilde{y}_i$ and confidence score $c_i = \text{conf}(x_i, I)$. To focus on training samples that offer high information gain, we enforce an *informative filter* that retains pairs $(x_i, \tilde{y}_i)$ with moderate confidence:

$$\tau_{\text{low}} \leq c_i \leq \tau_{\text{high}}, \tag{9}$$

where τ_{low} and τ_{high} are thresholds set to 0.25 and 0.75, respectively. This criterion effectively discards trivial samples where the model is already certain ($c_i > 0.75$) as well as highly unstable or noisy generations ($c_i < 0.25$). The final curated training set $\mathcal{S}$ is formed by collecting all retained pairs across images, up to a budgeted size, to optimize the Multimodal Reasoner via GRPO.

Per-Sample Verifiable Reward. For a question $x_i \in \mathcal{S}$ with pseudo-label $\tilde{y}_i$, the Multimodal Reasoner generates a group of G candidate answers $\{y_j\}_{j=1}^G$. Each sampled answer receives the binary reward

$$r_j = \begin{cases} 1, & \text{if } y_j = \tilde{y}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

These rewards are group-normalized to produce advantages $\hat{A}_j$ as in Eq. 2 (with the Reasoner's rewards), and S_ϕ is updated by minimizing $\mathcal{L}_{\text{GRPO}}(\phi)$ as in Eq. 3.

3. Experiments

3.1. Benchmarks and Evaluation Protocol

We use existing image datasets from Vision-47K [9, 21], which contains 47K web images collected from diverse domains, e.g. including charts, medical images, exams, textbooks, and driving simulations.[2] We train three backbone models using *VisPlay*—Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, and Mimo-7B-SFT.[3] We run evaluation across three multimodal domains [21].

- **General Visual Understanding.** We measure performance on four established benchmarks. MM-Vet [44] provides a unified LLM-based score across recognition, OCR, and visual math tasks. MMMU [45] evaluates cross-modal reasoning and subject knowledge through 11.5K college-level. RealWorldQA [38] contains roughly 700 real-world images paired with spatially grounded questions. VisNumBench [37] focuses on visual number sense, covering around 1.9K questions involving numerical attributes and estimation tasks.
- **Multimodal Mathematical Reasoning.** MathVerse [47] consists of 2.6K diagram-centric questions, provided in multiple visual-text formats. MATH-Vision [34] includes around 3K competition-level problems across 16 subjects and five difficulty tiers.
- **Visual Hallucination Detection.** HallusionBench [11] is used to analyze model errors, distinguishing between language-only hallucinations and visual-illusion errors, with a simple yes/no evaluation format.

3.2. Main Results

We use LLM-as-a-judge to assess the correctness of the answers to ensure more robust evaluation [10, 19]. We present the outputs of the Multimodal Reasoner and analyze its reasoning ability progression in Table 1. We summarize the main findings of our experimental results below.

- ***VisPlay* consistently improves overall performance across different models.** All models trained with *VisPlay* consistently surpass both their corresponding base

[2] We only use the images without the questions and answers. Details of the dataset breakdowns are in supplement materials.

[3] The detailed training configurations are provided in the supplementary material.

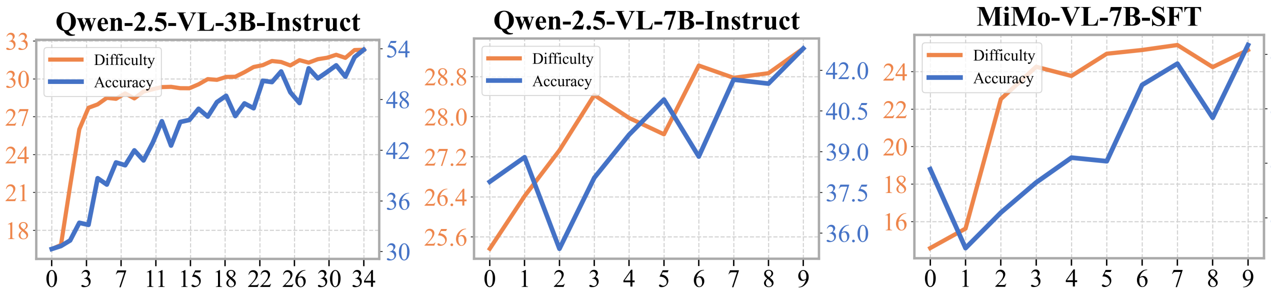


Figure 3. Changes in question difficulty (orange, left axis) and problem-solving accuracy (blue, right axis) during Image-Conditioned Questioner and Multimodal Reasoner training across three VLMs.

models and the Base Challenger over successive training iterations. Qwen2.5-VL-3B shows a remarkable improvement, with the average score increasing from 30.61 at baseline to 44.16 after the first iteration and reaching 47.27 at the third. Qwen2.5-VL-7B and MiMo-VL-7B follow similar upward trends, improving from 40.41 to 48.61 and 43.56 to 45.69, respectively. These results demonstrate the robust generalization ability and scalability of the proposed self-evolving framework across different models and model sizes.

- **Performance gains across diverse task types.** *VisPlay* shows improvements across general visual understanding tasks, visual reasoning or math benchmarks, and are more robust to hallucination. For Qwen2.5-VL-3B, the Hallucination score rises from 32.81 to 94.95 by the second iteration, showing a substantial enhancement in factual grounding. Similar patterns are observed in other models—reasoning benchmarks consistently improve without compromising accuracy on general understanding tasks—demonstrating that *VisPlay* effectively strengthens both task-specific reasoning and cross-domain multimodal generalization.
- **Iterative co-evolution between the Questioner and Reasoner drives improvement.** Performance trajectories across iterations highlight the co-evolution between the Questioner and Reasoner. As the Questioner generates more diverse and challenging queries, the Reasoner—trained with GRPO using high-quality silver supervision—learns to handle increasingly complex reasoning steps. This iterative loop allows both components to reinforce each other, leading to continual improvement in reasoning quality. The results indicate that the co-evolutionary design of *VisPlay* provides a scalable path toward self-improving multimodal intelligence.

3.3. Performance Comparison with Human-Annotated Data

We conduct a performance comparison between models trained with *VisPlay* and those trained using human-curated image–question–answer pairs from the Vision-47K dataset under standard GRPO for one epoch, as shown in Table 3 for Qwen2.5-VL-3B and 7B. Although this experiment is not an ablation study in the strict sense, it provides a clear view of how our fully automated training pipeline performs relative to conventional supervised training. Overall, we observe that models trained with *VisPlay* achieve competitive average accuracy compared with those trained on real, human-written data. While performance on several task categories differs slightly, the general trend indicates that the self-evolving process can produce training signals of sufficient quality to improve base VLMs capabilities. These findings suggest that even in settings where human annotations are costly, limited, or unavailable, our framework can still serve as an effective and scalable alternative, enabling VLMs to develop stronger generalization abilities without depending on manual supervision.

3.4. Co-Evolution Dynamics of Two Roles

- **The Evolution of Question Difficulty and Solution Accuracy.** To analyze the co-evolution dynamics of the two roles, we examine the changes in question difficulty (orange, left axis) and problem-solving accuracy (blue, right axis) across three VLMs during the first training iteration (Figure 3). Question difficulty is operationalized as the Reasoner's model-perceived difficulty, derived from the confidence score defined in Eq. 4 in Section 2.3. Across all models, a consistent co-evolution pattern emerges. The Image-Conditioned Questioner's difficulty curves exhibit a general upward trend, with initial increments followed by sustained growth, indicating the role's ability to progressively formulate more challenging visual questions. Concurrently, the Multimodal Reasoner's accuracy curves, despite minor fluctuations, show a complementary upward trajectory. This means that as question difficulty rises, the Reasoner adapts and enhances its problem-solving capability, with both metrics reinforcing each other's improvement over iterations. Such mutual reinforcement validates *VisPlay*'s core mechanism, where

Algorithm 1: *VisPlay*: Self-Evolving RL for Vision-Language Models

Input: Initial models Q_θ, S_ϕ; Image dataset $\mathcal{I}$; Group size G; Reasoner samples m; Dataset budget N; Thresholds $\tau_{\text{low}} = 0.25, \tau_{\text{high}} = 0.75$.

Output: Evolved models Q_θ and S_ϕ.

- **for** *each self-play iteration* **do**
  - **for** *each image batch* $I \in \mathcal{I}$ **do**
    - Sample question group $\{x_i\}_{i=1}^{G} \sim Q_\theta(\cdot \mid I)$;
    - **for** *each question* x_i **do**
      - Sample m answers $\{y_j\}_{j=1}^{m} \sim S_\phi(\cdot \mid I, x_i)$;
      - Compute confidence $c_i \leftarrow \text{conf}(x_i, I)$ via majority vote (Eq. 4);
      - Compute uncertainty reward $r_{\text{unc}} \leftarrow 1 - |2c_i - 1|$;
      - Compute diversity penalty r_{div} via clustering (Eq. 6);
      - Final reward: $r_i \leftarrow \mathbb{1}_{\text{valid}}(x_i) \cdot \text{ReLU}\big(r_{\text{unc}} - r_{\text{div}}\big)$;
    - **end**
    - Update Q_θ via GRPO using rewards $\{r_i\}_{i=1}^{G}$;
  - **end**
  - Initialize curated dataset $\mathcal{S} \leftarrow \emptyset$;
  - **for** *each image* $I \in \mathcal{I}$ **do**
    - Generate N candidate questions via Q_θ ;
    - **for** *each candidate* x_k **do**
      - Obtain pseudo-label $\tilde{y}_k$ and confidence c_k from S_ϕ;
      - **if** $\tau_{\text{low}} \leq c_k \leq \tau_{\text{high}}$ **then**
        - Add $(I, x_k, \tilde{y}_k)$ to $\mathcal{S}$;
      - **end**
    - **end**
  - **end**
  - **for** *each minibatch* $(I, x, \tilde{y}) \in \mathcal{S}$ **do**
    - Sample G answers $\{y_j\}_{j=1}^{G} \sim S_\phi(\cdot \mid I, x)$;
    - Compute binary rewards $r_j \leftarrow \mathbb{1}(y_j = \tilde{y})$;
    - Update S_ϕ via GRPO using rewards $\{r_j\}_{j=1}^{G}$;
  - **end**
- **end**

the two interacting roles drive scalable self-evolution in multimodal reasoning.

- **The Evolution of Capabilities and Data Accuracy.** Building on the interaction between question difficulty and solution accuracy, we further analyze how the Reasoner's capabilities and the quality of pseudo-labeled data evolve across iterations. For each training iteration, the Questioner generates questions for the same 200 images, and Reasoners from each iteration attempt to answer them. As shown in Table 2, the Reasoner's accuracy steadily improves across iterations (e.g., from 44.0 to 49.0 on first-iteration questions), while the estimated accuracy of pseudo-labels slightly declines (from 72 to 61), reflecting increasing question difficulty. These trends highlight the co-evolution of model reasoning ability and data complexity during self-improving training.

3.5. Case Study on Question Difficulty Evolution

Table 4 presents example questions generated by the self-evolving Vision-Language Models across three training iterations. Iteration 1 questions focus on direct observation, such as counting or identifying objects. Iteration 2 introduces relational and comparative reasoning, requiring the model to assess differences or evaluate spatial angles. Iteration 3 further increases complexity with multi-step reasoning and inference, including precise localization and causal relationships. This progression demonstrates a systematic increase in question difficulty, providing increasingly challenging training signals that encourage the model to adapt and improve its reasoning capabilities. Such a design ensures that both the Questioner and Reasoner co-evolve, progressively enhancing the overall performance of the system.

4. Related Work

Post-Training for Vision-Language Models Recent research in post-training of vision-language models (VLMs) has shifted from supervised fine-tuning (SFT) toward reinforcement learning (RL) paradigms. Earlier works such as LLaVA [14, 25] primarily rely on SFT to align a language model backbone with a visual information via a projection layer, enabling multimodal instruction-following and visual reasoning. However, as base model quality improves, RL-based post-training has emerged as a more powerful alternative. In particular, R1-style training [7] has gained attention for its ability to enhance visual reasoning. A key insight behind this success is that RL becomes effective only when the base model is sufficiently capable to self-explore reasoning trajectories [1, 5, 17, 18]. Despite these advances, most existing RL-based VLM approaches still depend on annotated multimodal datasets, which are costly and difficult to scale [8, 9, 12, 20, 23, 28, 46, 48, 49]. To reduce reliance on human supervision, several recent works explore VLM self-play paradigms in games. Vision-Zero [35] and Game-RL [33] train VLMs with simulated game data to improve their general reasoning ability. Nonetheless, these methods often continue to depend on external models or tools for training data generation.

Self-Evolving In Large Language Models Recent work has focused on enabling large language models (LLMs) to

Table 1. Comprehensive results on visual reasoning benchmarks. Each base model is evaluated against two settings: a *VisPlay* (❄ challenger) baseline, in which the Reasoner is trained on questions produced by an untrained Challenger, and our iterative *VisPlay* framework. The highest performance reached during training for each model is emphasized in bold. We take accuracy as the metric.

	General Visual Understanding				**Visual Math**		**Hallucination**	
Methods	**MMMU**	**MM -Vet**	**RealWorld QA**	**VisNum Bench**	**Math Verse**	**MATH -Vision**	**Hallusion Bench**	**Avg.**
Qwen2.5-VL-3B-Instruct								
Base Model	19.95	36.24	49.28	27.08	26.14	20.23	32.81	30.61
VisPlay (❄ challenger)	23.34	43.58	57.78	29.33	33.50	23.39	64.88	33.77
VisPlay (Iter 1)	29.40	**48.62**	67.06	30.01	29.67	22.57	91.80	44.16
VisPlay (Iter 2)	33.37	44.50	65.62	29.64	32.36	24.67	**94.95**	44.87
VisPlay (Iter 3)	**37.11**	38.07	**71.90**	**39.15**	**35.15**	**29.97**	90.54	**47.27**
Qwen2.5-VL-7B-Instruct								
Base Model	23.10	44.95	57.52	32.57	33.78	24.05	66.88	40.41
VisPlay (❄ challenger)	35.24	45.87	69.67	32.41	35.13	26.22	78.13	38.33
VisPlay (Iter 1)	28.94	46.33	62.61	28.65	33.88	26.91	80.34	44.53
VisPlay (Iter 2)	27.07	42.66	60.92	27.08	36.32	25.00	67.72	40.97
VisPlay (Iter 3)	**38.27**	**46.33**	**69.67**	**32.57**	**39.14**	**31.15**	**92.32**	**48.61**
MiMo-VL-7B-SFT								
Base Model	30.22	**59.17**	78.17	44.80	41.80	25.33	87.17	43.56
VisPlay (❄ challenger)	27.54	58.72	63.27	49.66	38.78	24.80	57.83	39.63
VisPlay (Iter 1)	25.67	56.42	69.54	51.59	40.20	25.13	**87.59**	43.16
VisPlay (Iter 2)	**34.07**	55.96	**78.69**	51.18	41.65	28.45	86.65	45.58
VisPlay (Iter 3)	28.24	56.88	71.50	**52.69**	**46.02**	**29.44**	74.55	**45.69**

Table 2. Analysis of model performance and data quality. The shaded column indicates the estimated accuracy of the self-generated pseudo-labels for each question set, as determined using ChatGLM-Flash.

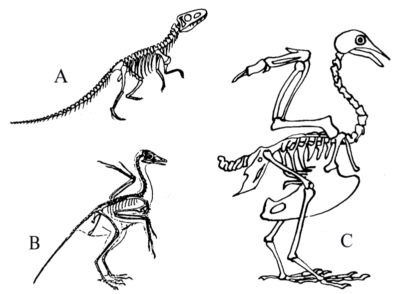
	Performance of Evaluated Model (vs. Ground Truth)				
	Base Model	Reasoner (Iter 1)	Reasoner (Iter 2)	Reasoner (Iter 3)	Pseudo-Label Acc.
$\mathcal{D}_{\text{Iter 1}}$	39.0	44.0	45.5	49.0	72.0
$\mathcal{D}_{\text{Iter 2}}$	37.5	42.5	44.0	47.5	65.0
$\mathcal{D}_{\text{Iter 3}}$	36.0	40.0	41.5	45.0	61.0

Table 3. Performance comparison between *VisPlay* and standard GRPO training with human-labeled data. Although *VisPlay* relies entirely on self-generated supervision, it achieves competitive overall accuracy and significantly reduces hallucination. This demonstrates that our self-evolving framework can meaningfully enhance VLM performance even in the absence of human-annotated datasets.

	General Visual Understanding				**Visual Math**		**Hallucination**	
Methods	**MMMU**	**MM -Vet**	**RealWorld QA**	**VisNum Bench**	**Math Verse**	**MATH -Vision**	**Hallusion Bench**	**Avg.**
Qwen2.5-VL-3B-Instruct								
Standard GRPO	40.3	49.5	63.0	36.7	42.8	29.9	67.4	47.1
VisPlay (Iter 3)	37.1	38.1	71.9	39.2	35.2	30.0	90.5	47.3
Qwen2.5-VL-7B-Instruct								
Standard GRPO	39.8	51.8	66.6	43	53.2	33.8	66.6	50.7
VisPlay (Iter 3)	38.3	46.3	69.7	32.6	39.1	31.2	92.3	48.6

self-evolve their reasoning capabilities with minimal to zero human supervision. Various approaches have been proposed to achieve this. General unsupervised self-training frameworks like Genius [41] and Deep Self-Evolving Reasoning [26] aim for advanced reasoning. A common theme is data-free training or starting from zero data, as ex-

Table 4. Examples of challenging questions generated by the self-evolving Vision-Language model across three training iterations. The questions progressively increase in complexity, illustrating the growth in difficulty of the Questioner's outputs over iterations. Images on the left and right correspond to the visual context for each question. The questions are raised by Qwen2.5-VL-3B-Instruct.

Challenging Examples from Self-Evolving Trained Vision-Language Model		
		
Question (Iter 1)	Approximately how many lung fields are visible in the X-ray image?	Which skeletal structure most likely belongs to a bird with hollow bones?
Question (Iter 2)	On a thoracic x-ray, the right lobe of the lung is more spread out compared to the left lobe. If the right lobe is given a score of 1 and the left lobe is given a score of 0, what is the difference in scores between the right and left lung lobes?	On which figure does the long neck of the dinosaur have the greatest horizontal angle with the vertical axis?
Question (Iter 3)	On which rib is the line approximately 2.5 cm above the image's midpoint?	Which skeletal structure is most likely to have evolved secondary to flying abilities and which is less likely to have this trait?

plored in R-Zero [13], Absolute zero [50], and Language self-play [15]. Other methods adapt self-play for specific goals or environments. For example, SPICE [24] improves reasoning in corpus environments, Search Self-play [27] pushes capabilities via search, and SPELL [43] focuses on evolving long-context models. Additionally, some methods utilize interactions between multiple agents to collectively bootstrap reasoning abilities, as seen in Socratic-Zero [36] and Multi-Agent Evolve [4].

5. Limitation

While our work introduces a scalable, self-evolving framework, we acknowledge two primary limitations that suggest directions for future research. First, due to computational constraints, our experiments were limited to the Qwen2.5-VL and MiMo-VL families. The scalability and effectiveness of *VisPlay* on significantly larger VLMs (e.g., $\geq$ 10B parameters) is still an important open question. Second, our framework lacks a definitive verification method for the self-generated data. While our GRPO policy indirectly optimizes for quality, developing more robust, automated methods to verify data faithfulness and prevent error accumulation is a key area for future investigation.

6. Conclusion

We present *VisPlay*, a self-evolving RL framework that enables vision-language models to autonomously improve from unlabeled images. By decomposing a VLM into an Image-Conditioned Questioner and a Multimodal Reasoner and optimizing them via GRPO, our method balances challenge and accuracy without human supervision. Experiments show consistent gains in reasoning, compositional generalization, and hallucination reduction across multiple benchmarks. *VisPlay* demonstrates that scalable, self-improving multimodal intelligence is achievable. By iteratively generating and learning from its own experiences, a model can refine its capabilities beyond human-labeled data. This framework opens avenues for richer multimodal interactions and cross-domain adaptation, pointing toward intelligence systems that can continually evolve autonomously. Our results suggest a promising path toward truly autonomous vision-language systems that improve themselves over time.

Acknowlegement

This research was partially supported by the NExT++ Research Centre and ST Engineering under a collaborative research project (Agreement No. 2024-01813), and by the Fundamental Research Funds for the Central Universities of China (Grant No. PA2025IISL0106).

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 6

[2] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. An introduction to vision-language modeling, 2024. 1

[3] Xiuwei Chen, Wentao Hu, Hanhui Li, Jun Zhou, Zisheng Chen, Meng Cao, Yihan Zeng, Kui Zhang, Yu-Jie Yuan, Jianhua Han, Hang Xu, and Xiaodan Liang. C2-evo: Co-evolving multimodal data and model for self-improving reasoning, 2025. 1

[4] Yixing Chen, Yiding Wang, Siqi Zhu, Haofei Yu, Tao Feng, Muhan Zhang, Mostofa Patwary, and Jiaxuan You. Multi-agent evolve: Llm self-improve through co-evolution. *arXiv preprint arXiv:2510.23595*, 2025. 8

[5] Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. Visionllama: A unified llama backbone for vision tasks, 2024. 6

[6] Jeff Clune. Ai-gas: Ai-generating algorithms, an alternate paradigm for producing general artificial intelligence. *arXiv preprint arXiv:1905.10985*, 2019. 1

[7] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 6

[8] Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward, 2025. 6

[9] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms, 2025. 4, 6

[10] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. 4

[11] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385, 2024. 4

[12] Yicheng He, Zheng Zhao, Zhou Kaiyu, Bryan Dai, Jie Fu, and Yonghui Yang. Codecircuit: Toward inferring llm-generated code correctness via attribution graphs. *arXiv preprint arXiv:2602.07080*, 2026. 6

[13] Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. 2025. 1, 8

[14] Mengzhao Jia, Zhihan Zhang, Wenhao Yu, Fangkai Jiao, and Meng Jiang. Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training, 2024. 6

[15] Jakub Grudzien Kuba, Mengting Gu, Qi Ma, Yuandong Tian, and Vijai Mohan. Language self-play for data-free training. *arXiv preprint arXiv:2509.07414*, 2025. 8

[16] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. 2

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 6

[18] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pages 1587–1606, 2025. 1, 6

[19] Zongxia Li, Xiyang Wu, Ishani Mondal, Alexa Siu, Jordan Lee Boyd-Graber, and Ani Nenkova. LLM-as-a-judge failures at automating the identification of poor quality outputs in free-form texts. In *Proceedings of The 5th New Frontiers in Summarization Workshop*, pages 1–16, Hybrid, 2025. Association for Computational Linguistics. 4

[20] Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding, 2025. 6

[21] Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, et al. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*, 2025. 4

[22] Zongxia Li, Hongyang Du, Chengsong Huang, Xiyang Wu, Lantao Yu, Yicheng He, Jing Xie, Xiaomin Wu, Zhichao Liu, Jiarui Zhang, et al. Mm-zero: Self-evolving multi-model vision language models from zero data. *arXiv preprint arXiv:2603.09206*, 2026. 1

[23] Zongxia Li, Hongyang Du, Dawei Liu, Xiyang Wu, Lantao Yu, Jingxi Chen, Fuxiao Liu, Xiaomin Wu, Jing Xie, Chengsong Huang, Yicheng He, and Guangyao Shi. Multimodal video generation models with audio: Present and future. *https://doi.org/10.13140/RG.2.2.26531.31528*, 2026. 6

[24] Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. Spice: Self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*, 2025. 1, 8

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 6

[26] Zihan Liu, Shun Zheng, Xumeng Wen, Yang Wang, Jiang Bian, and Mao Yang. Deep self-evolving reasoning. *arXiv preprint arXiv:2510.17498*, 2025. 7

[27] 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. *arXiv preprint arXiv:2510.18821*, 2025. 8

[28] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl, 2025. 6

[29] Jürgen Schmidhuber. Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial general intelligence*, pages 199–226. Springer, 2007. 1

[30] Atharva Sehgal, Patrick Yuan, Ziniu Hu, Yisong Yue, Jennifer J. Sun, and Swarat Chaudhuri. Self-evolving visual concept library using vision-language critics, 2025. 1

[31] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2

[32] Qwen Team. Qwen2.5-vl, 2025. 1, 2

[33] Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, Chaoran Tao, Zhiyuan Guo, Jizhou Yu, Tianhao Cheng, Zhiheng Xi, Changhao Jiang, Zhangyue Yin, Yining Zheng, Weifeng Ge, Guanhua Chen, Tao Gui, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. Game-rl: Synthesizing multimodal verifiable game data to boost vlms' general reasoning, 2025. 6

[34] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. 4

[35] 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-improvement via strategic gamified self-play, 2025. 6

[36] Shaobo Wang, Zhengbo Jiao, Zifan Zhang, Yilang Peng, Xu Ze, Boyu Yang, Wei Wang, Hu Wei, and Linfeng Zhang. Socratic-zero: Bootstrapping reasoning via data-free agent co-evolution. *arXiv preprint arXiv:2509.24726*, 2025. 8

[37] Tengjin Weng, Jingyi Wang, Wenhao Jiang, and Zhong Ming. Visnumbench: Evaluating number sense of multimodal large language models. *arXiv preprint arXiv:2503.14939*, 2025. 4

[38] xAI. Realworldqa: Real-world spatial understanding benchmark. `https://x.ai/blog/grok-1.5v-and-realworldqa`, 2024. CC BY-ND 4.0 license. Benchmark dataset released with Grok-1.5 Vision. 4

[39] Zhishang Xiang, Chengyi Yang, Zerui Chen, Zhimin Wei, Yunbo Tang, Zongpei Teng, Zexi Peng, Zongxia Li, Chengsong Huang, Yicheng He, et al. A systematic survey of self-evolving agents: From model-centric to environment-driven co-evolution. 1

[40] LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. 2

[41] Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Qiushi Sun, Kanzhi Cheng, Junxian He, Jun Liu, and Zhiyong

Wu. Genius: A generalizable and purely unsupervised self-training framework for advanced reasoning. *arXiv preprint arXiv:2504.08672*, 2025. 7

[42] Yuqing Yang, Yan Ma, and Pengfei Liu. Weak-to-strong reasoning. *arXiv preprint arXiv:2407.13647*, 2024. 1

[43] Ziyi Yang, Weizhou Shen, Ruijun Chen, Chenliang Li, Fanqi Wan, Ming Yan, Xiaojun Quan, and Fei Huang. Spell: Self-play reinforcement learning for evolving long-context language models. *arXiv preprint arXiv:2509.23863*, 2025. 8

[44] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 4

[45] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 4

[46] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization, 2025. 6

[47] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024. 4

[48] Xiaoyan Zhang and Jiangpeng He. One adapter for all: Towards unified representation in step-imbalanced class-incremental learning. *arXiv preprint arXiv:2603.10237*, 2026. 6

[49] Xiaoyan Zhang, Zechen Bai, Haofan Wang, and Yiren Song. Sigma: Selective-interleaved generation with multi-attribute tokens. *arXiv preprint arXiv:2602.07564*, 2026. 6

[50] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025. 1, 8