VisPlay: Self-Evolving Vision-Language Models

Supplementary Material

A.1 Detailed Training Dataset and Benchmarks

Training Dataset We use the image data from Vision-47K dataset [12, 25], which contains 47,000 web-sourced images covering a wide variety of domains. The dataset includes charts, medical images, educational exams, textbook illustrations, and driving simulation frames. For our purposes, we exclusively use the images themselves, omitting any associated questions and answers. The dataset consists of approximately 10K charts, 8K medical images, 12K educational images (from exams and textbooks), 7K driving scenes, and 10K miscellaneous images from various domains. All images were standardized to a resolution of 224×224 pixels for model training.

Backbone Models and Training We trained three backbone models using *VisPlay*:

- **Qwen2.5-VL-3B-Instruct** : 3 billion parameters, fine-tuned with multimodal instruction data to enhance reasoning over visual-text tasks.
- **Qwen2.5-VL-7B-Instruct** : 7 billion parameters, trained under the same protocol with extended batch sizes and longer training schedules to improve complex reasoning and generalization.
- **Mimo-VL-7B-SFT** : 7 billion parameters, optimized with supervised fine-tuning on multimodal datasets for better alignment with human instructions.

General Visual Understanding Four established benchmarks are used:

- **MM-Vet** [47]: Evaluates recognition, OCR, and visual math abilities using a unified LLM-based scoring metric. The dataset contains over 5,000 test samples with detailed scoring for each subtask.
- **MMMU** [48]: Cross-modal reasoning benchmark with 11.5K college-level multiple-choice questions spanning six academic disciplines. Each question is image-based and designed to test subject knowledge and reasoning ability.
- **RealWorldQA** [41]: Contains approximately 700 real-world images paired with spatially grounded questions. Evaluation emphasizes spatial reasoning and contextual understanding.
- **VisNumBench** [40]: Focused on visual number sense, includes roughly 1.9K questions involving numerical attributes, comparisons, and estimations.

Multimodal Mathematical Reasoning Two specialized benchmarks:

- **MathVerse** [50]: 2.6K diagram-centric questions covering geometry, functions, and algebra, provided in multiple visual-text formats.
- **MATH-Vision** [37]: Approximately 3K competition-level problems across 16 subjects and five difficulty tiers. Focuses on integrating visual information into advanced mathematical reasoning.

Visual Hallucination Detection **HallusionBench** [14] is used to identify model errors caused by either language-only hallucinations or visual illusions. Evaluation is conducted in a simple yes/no format, enabling precise measurement of hallucination rates and error types.

A.2 Training Configuration

Image-based Questioner Configuration. The Questioner is trained using a vision–language model with a maximum context window of 8192 tokens. The training set consists of 47K multimodal samples (Vision-SR1-47K), and evaluation is performed on the MMStar benchmark. Each sample uses `problem`, `answer`, and `images` as the prompt, label, and image fields, respectively. During rollouts, the Questioner generates 8 candidate questions per input. For the 3B model, we train for 20 steps, and for the 7B model, we train for 10 steps. In this setup, four GPUs run the vLLM service to provide reward signals, while the other four GPUs are used for training. The model parameters are loaded from a specified Questioner checkpoint, and all checkpoints are saved under the designated experiment directory. Validation before training is disabled to maximize efficiency during early-stage learning.

Multimodal Reasoner Configuration. The Solver is trained using chain-of-thought reinforcement learning. Its output length is capped at 4096 tokens, and prompts are constructed using a dedicated Jinja template to enforce a consistent reasoning format. Training uses the self-play dataset produced by the Questioner, while evaluation again uses MMStar. To ensure stable learning under long sequences, we adopt a conservative micro-batch size of 1 for both updates and experience rollouts. The rollout engine supports up to 20K batched tokens per forward pass. The number of training steps for the Solver is set to be the same as for the Questioner.

***Image-Conditioned Questioner* Prompt Template**

You are an intelligent Question Generator. Your task is to create a question based on the given image.
Requirements (must follow exactly):
1. Analyze the image carefully and understand all details. 2. Generate exactly one question that is directly related to the image. 3. Choose the question type from only one of the following: - multiple choice (Yes/No or four options labeled A, B, C, D; only one correct answer) - numerical (requires a specific numeric answer) - regression (requires predicting a continuous value, such as a measurement, quantity, or coordinate) 4. The question must require analysis or reasoning, not just description. 5. Output must be strictly in format $< question > X < /question >$, with nothing else:
Strict rules: - Do not use any other labels, punctuation, or formatting. - Do not add commentary, explanations, or extra text. Example of correct output:
$< question > How\ \ many\ \ clubs\ \ are\ \ there < /question >$

***Multimodal Reasoner* Prompt Template**

Please reason step by step carefully based on the question: + content + and the image. After completing your reasoning, you MUST output the final, clean, and concise answer strictly inside + \boxed{} + . The final answer MUST appear inside \boxed{}, and nowhere else. If there is no boxed answer, your response is considered incorrect.

***LLM-as-Judge* Prompt Template**

You are an answer evaluation assistant. Your task is to judge whether two answers are substantially equivalent. When evaluating, you should ignore superficial differences such as format, spaces, punctuation, case, etc., and focus on whether they are consistent in core content, logical meaning and information expression. The judgment criteria should be lenient and inclusive, as long as the expressed meaning is basically the same, it is considered equivalent.

A.3 Prompt Templates

The following three prompt templates define the core interaction structure used in our self-evolving Vision-Language Model. In this setup, multiple specialized roles—such as a question generator, a multimodal reasoner, and an evaluator—are orchestrated to form an autonomous learning loop. Each template specifies precise behavioral constraints that allow the model to generate tasks, solve them with step-by-step reasoning, and assess answer consistency. Together, these components establish a controlled self-play environment that enables the model to iteratively refine its reasoning capabilities without relying on external supervision.