A Multi-Model Self-Evolving Framework for Zero-Data Document Understanding via Axiomatic Synthetic Refinement

Yuwen Tao Fuxiao Liu

Abstract

The scalability of Vision-Language Models (VLMs) for structured data understanding is fundamentally bottlenecked by the scarcity of high-quality, human-curated datasets. We present Omni-Zero-Doc, a self-evolving framework designed for zero-resource document intelligence. By orchestrating a tri-role agentic system—Proposer, Coder, and Solver—integrated with the Style-Isomorphic Evidence-Grounded Evolution (SIEGE) engine, we enable models to programmatically synthesize high-fidelity, structure-aware visual data. To ensure logical consistency, we employ Group Relative Policy Optimization (GRPO) coupled with a multi-stage verifiable reward mechanism and axiomatic filtering to eliminate hallucinations. Across benchmarks including DocVQA, MathVista, and MMMU, Omni-Zero-Doc achieves a +14.0% ANLS gain, demonstrating that iterative self-play can sustain performance without external supervision.

1. Introduction

Self-evolving Mechanisms (Xiang et al.; Schmidhuber, 2007; Clune, 2019; Huang et al., 2025a; Yue et al., 2026) represent a promising frontier for advancing foundational models by allowing them to learn from independently generated experiences. While Large Language Models (LLMs) have achieved zero-data improvement through task generation and code execution, VLMs are traditionally constrained by the "Visual Modality Gap," requiring seed images to bootstrap the evolution process.

Existing VLM self-evolution approaches often remain conditioned on pre-existing static image datasets (He et al., 2025; Thawakar et al., 2025; Wang et al., 2026), shifting the bottleneck to image collection. Furthermore, VLMs (Liu et al., 2023c; Li et al., 2025; Huang et al., 2025b; Li et al., 2025; Feng et al., 2025; Liu et al., 2023b; Fei et al., 2024) frequently encounter "Structural Blindness" when applied to complex document layouts, such as financial reports or academic manuscripts. This is driven by resolution constraints and a lack of data requiring multi-step logical deduction. We introduce **Omni-Zero-Doc**, the first tri-role RL framework for multimodal zero-data evolution. By leveraging a generative Coder to render images from code and a SIEGE engine to refine instructions through axiomatic filtering, we establish a scalable path toward self-improving document intelligence (Xu et al., 2024)

2. The Omni-Zero-Doc Methodology

2.1. Tri-Role Interaction Architecture

Omni-Zero-Doc expands the traditional dual-role setup into a specialized tri-role interaction pipeline, where all roles are initialized from the same base model: **Proposer (π_P):** Formulates abstract visual concepts, textual descriptions c, and paired questions (q_{easy}, q_{hard}). It aims to generate "Goldilocks tasks" that challenge the solver without being impossible. **Coder (π_D):** Acts as the visual synthesis engine. It translates concepts into executable code (Python/SVG) to render images, ensuring absolute data accuracy where every pixel corresponds precisely to underlying values. **Solver (π_S):** Engages with the synthesized content to perform multi-step multimodal reasoning. It utilizes a Chain-of-Thought (CoT) trace, thinking step-by-step within `<think>` tags.

2.2. Reinforcement Learning with Verifiable Rewards

We employ Group Relative Policy Optimization (GRPO) to update each role sequentially. For a prompt p, the advantage for response x_i is:

$$\hat{A}_i = \frac{r_i - mean(r_1, \ldots, r_N)}{std(r_1, \ldots, r_N) + \epsilon_{norm}} \tag{1}$$

The policy is updated by maximizing a clipped surrogate objective. The Proposer's reward R_P balances solvability, difficulty, and diversity:

$$R_p = \frac{1}{N} \sum (\mathbb{I}_{exec} \cdot (\min(R_{solv}, 0.5) + R_{diff})) + r_{div} \tag{2}$$

where R_{solv} is verified by the solver on q_{easy}, and R_{diff} peaks when the solver is maximally uncertain.

3. SIEGE: Axiomatic Data Refinement

To prevent "hallucination propagation," the SIEGE engine applies four Refinement Axioms: **Evidence Grounding Axiom:** Prunes samples where the model introduces external knowledge not present in the rendered image. **Brevity and Format Control:** Filters out redundant transition tokens (e.g., "monotonically") that lead to "Attention Drift". **Syntactic Isomorphism:** Enforces strict alignment with the length and complexity of expert inquiry patterns. **Ambiguity Filtering:** Excludes queries demanding precise scalar values from low-resolution charts.

The construction of our synthetic training corpus, **Omni-Doc-Zero**, follows a hierarchical "Conception-to-Rendering" pipeline that requires zero human-annotated seeds. Unlike traditional datasets that rely on static image crawling, our engine synthesizes data across seven major categories: `data_chart`, `diagram`, `geometry`, `timeline`, `map`, `table`, and `other`. **Abstract Visual Prompting:** The Proposer is prompted to act as an expert Visual Content Designer, generating high-level captions (c) that specify complex attributes such as multiple data series, annotations, and non-trivial geometric patterns. **Programmatic Rendering:** The Coder translates these captions into executable SVG or Python code. To ensure "Layout-Logic" fidelity, the code generation is strictly grounded in numerical precision, where every graphical element (e.g., tick marks, bar heights) is deterministically mapped to the intended data values. **QA Pair Synthesis:** For each rendered scene, the Proposer generates an *Easy Question* (q_{easy}) for verifying basic visual recognition and a *Hard Question* (q_{hard}) necessitating multi-step logical deduction.

To mitigate the risk of "Model Collapse" and hallucination propagation, we implement the **Style-Isomorphic Evidence-Grounded Evolution (SIEGE)** framework for automated data auditing. **Axiomatic Filtering:** Each candidate sample (I, q, a) must pass a cascade of four logical axioms verified by a high-capacity heuristic judge (e.g., GPT-4o): **Evidence Grounding Axiom:** Prunes instances where the answer relies on linguistic priors or external knowledge rather than the synthesized pixels. **Brevity and Attention Drift Control:** Prunes responses with excessive verbosity or "wait" tokens, which are correlated with parsing failures. **Syntactic Isomorphism:** Ensures the linguistic complexity and professional tone of the instructions remain consistent with academic and financial reporting standards. **Ambiguity and Precision Axiom:** Excludes ill-posed queries, such as demanding exact scalar values from low-resolution or partially obscured charts.

Through this iterative refinement, we curated the **DocInstruct-100K** corpus. The evolution of data quality is evidenced by the step-wise retention rates shown in Table 1.

Filter Criterion	Initial → Filtered	Ret. (%)
Visual Grounding & Clarity	250k → 200k	80.0%
Brevity & Length Control	200k → 150k	75.0%
Redundant Token Pruning	150k → 125k	83.3%
Quality Consensus	125k → 100k	80.0%
Final Dataset	**100,000**	**40.0%**

Table 1. Step-wise data refinement statistics for Omni-Doc-Zero.

3.1. Difficulty Balancing: The Goldilocks Principle

A key feature of our dataset is the enforcement of the **Goldilocks Principle** during generation. We filter training data to retain only those examples where the Coder's rendering success rate is between 0.25 and 0.75, and the Solver's hard-question accuracy falls within the same range. This ensures the model is trained on samples that are neither trivially solvable nor impossibly difficult, maximizing the gradient signal for self-evolution.

4. Experiments

4.1. Experimental Setup

Model Configurations: We employ a range of base models to validate the scalability of our framework, including Qwen3-VL-Instruct (4B and 8B variants) and Mimo-VL-7B-SFT. All models are initialized with their respective base weights and undergo full fine-tuning with a trainable vision tower. **Training Procedure:** The self-evolution process is conducted in iterative loops. Each iteration consists of sequential training phases for the Proposer, Coder, and Solver, with each role trained for 20 steps, totaling 60 steps per iteration. We utilize a cluster of 8x RTX 6000 (96GB) or Nvidia Tesla V100 GPUs . For the document-specific fine-tuning, we employ a two-stage paradigm: Stage 1 focuses on modality alignment of document-centric visual tokens, and Stage 2 performs instruction tuning on the synthesized DocInstruct-100K dataset. **Hyper-parameters:** We set the learning rate to 1×10^{-6} with a weight decay of 1×10^{-4}. For reinforcement learning, we use a rollout group size of $N = 4$ for the Proposer and Coder, and $K = 5$ or 8 for the Solver's self-consistency voting.

4.2. Evaluation Benchmarks

Our evaluation suite covers four critical dimensions of multimodal intelligence: **General Visual Reasoning:** MMMU, MMMU-Pro, ChartQA, and MM-Vet, testing chart interpretation and object recognition. **Mathematical Reasoning:** MathVerse, MathVision, MathVista, and VisNumBench, assessing rigorous geometric and quantitative logic over visual inputs. **Document Understanding:** DocVQA, TextVQA, and InfographicVQA for single-page analysis;

MP-DocVQA and HDR for multi-page and hierarchical reasoning.**Hallucination Detection:** HallusionBench (Liu et al., 2023a) and MMSI, evaluating visual grounding versus linguistic priors.

4.3. Main Results and Comparative Analysis

The results in Table 1 demonstrate that Omni-Zero-Doc consistently outperforms base models and traditional baselines.

Zero-Data Evolution Gains: For the 8B model, we observe a steady improvement from 50.7% to 54.1% average accuracy across three iterations without any human data. The gains are most pronounced in visual math and document tasks, where logical depth is paramount. In DocVQA, our model reaches an ANLS of 72.5, a +14.0% leap over the LLaVA-1.5 baseline (58.5).

Scaling Beyond Iteration 3: We continued training the 8B model for two additional iterations. The average benchmark scores increased monotonically, reaching 56.6% at Iteration 5, suggesting that the self-evolution ceiling has not yet been reached.

4.4. Data Purity and Success Rate Analysis

The effectiveness of our Coder and the SIEGE refinement engine is further validated by internal metrics. **Coder Rendering Success:** As shown in our tracking, the 8B model's render success rate starts at ∼80% and converges toward 100% by Iteration 3. In contrast, the 4B model exhibits a lower success rate (∼60-78%), explaining its more modest performance gains. **Image Solvability:** We measure the percentage of successfully rendered images that allow the solver to answer q_{easy} correctly. This metric rises from ∼75% to over 85% for the 8B model, indicating that the synthesized visuals become increasingly faithful to the Proposer's intent. **Axiomatic Retention:** Our SIEGE engine filters ∼250k initial samples down to 100k high-purity pairs. The grounding filter and brevity control have the highest impact, with retention rates of 80.0% and 75.0% respectively.

4.5. Quantitative Evolution Analysis

Evolution of Code Generation Capability: A critical component of Omni-Zero-Doc is the Coder's ability to transform abstract descriptions into executable visual content. We track the *Coder Render Success Rate* and *Rendered Image Solvability* across 60 training steps.

As illustrated in Figure 1, all models exhibit a monotonic increase in rendering success. The Qwen3-VL-8B model, in particular, achieves nearly 100% rendering success by step 60. More importantly, the gap between rendering success and solvability narrows over time, suggesting that the model is not merely learning to generate "syntactically correct" code, but "semantically faithful" code that contains the necessary visual information to answer reasoning questions.

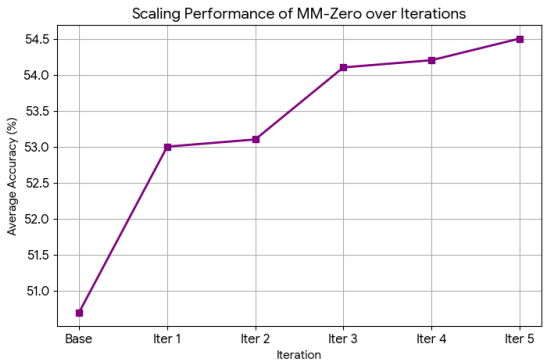

Figure 1. Evolution of Coder capabilities over training steps. The solid lines represent the Render Success Rate, while dashed lines indicate Image Solvability across three base models (4B, 7B, and 8B).

Multi-Generation Scaling Trends: To investigate the potential ceiling of self-evolution, we extended the training of the Qwen3-VL-8B model beyond the initial three iterations to a total of five iterations.

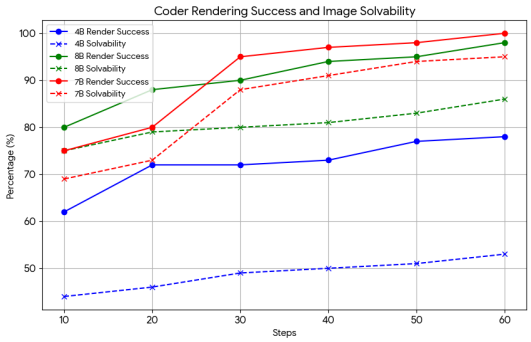

Figure 2. Scaling performance of Omni-Zero-Doc over five iterations on the Qwen3-VL-8B model. The Y-axis represents the average score across all evaluation benchmarks.

Figure 2 presents the performance trajectory. The average benchmark score increases from 50.7% (Base) to 54.1% (Iter 3) and reaches 54.5% at Iteration 5. While the rate of improvement slightly decelerates in later stages, the continued growth indicates that the tri-role interaction and axiomatic filtering effectively prevent the "model collapse" often observed in recursive self-training on unfiltered synthetic data.

4.6. Fine-grained Refinement Statistics

The effectiveness of the SIEGE engine's axiomatic filtering is demonstrated by the data retention rates during the

Model Variant	Iteration	DocVQA	MathVista	MMMU	ChartQA	MMSI	Avg.
Qwen3-VL-4B	Base	–	64.0	41.8	79.6	72.3	50.2
	Iter 1	–	68.5	44.5	83.4	73.2	53.5
	Iter 3	–	65.7	45.3	83.0	74.2	53.4
Qwen3-VL-8B	Base	61.2	67.7	46.6	76.9	72.8	50.7
	Iter 1	64.5	67.7	49.3	78.9	72.1	53.0
	Iter 3	**72.5**	67.2	**53.0**	**79.6**	**74.1**	**54.1**
Mimo-VL-7B	Iter 2	–	73.3	48.7	85.0	70.7	56.1

Table 2. Comprehensive results across General, Math, and Document benchmarks. ANLS is used for DocVQA.

synthesis of DocInstruct-100K. Starting from an initial pool of 250,000 raw samples, the *Visual Grounding & Clarity* filter removes 20% of samples due to unreliability. The *Brevity & Length Control* further prunes 25% to ensure concise and logic-dense responses. Finally, the *GPT-4o Quality Consensus* ensures a 40% overall retention rate, yielding 100,000 high-purity samples that are essential for achieving the observed +14.0% ANLS gain on DocVQA.

4.7. Error Analysis and Limitations

Despite the success, a qualitative error analysis reveals that the 4B model's limited gains are tied to its baseline inability to render complex multi-panel diagrams, leading to a "poverty of experience" during self-play. Future work should focus on curriculum-based rendering tasks to bootstrap smaller models more effectively.

4.8. Detailed Ablation Studies

We conduct a step-wise ablation on the SIEGE components and reward mechanisms to understand their contributions. As shown in Table 2, the transition from raw instruction pairs to CoT-structured responses provides a +3.6 ANLS gain. The grounding filter further adds +2.3, while brevity control (pruning "Attention Drift" tokens) adds +1.4. **Solvability Capping:** Removing the 0.5 cap on R_{solv} results in only a 2.3% improvement versus our 3.9% gain. Without the cap, the model hacks the reward by generating trivially easy captions to ensure solvability at the expense of difficulty. **Content Diversity:** Omitting the r_{ct} reward causes performance to decline after Iteration 1 (dropping to 49.4% by Iter 3). The model overfits to a narrow subset of "easy-to-render" visual types like simple histograms.

Configuration	ANLS (DocVQA)
Raw Synthesis (No Filtering)	64.2
+ CoT Reasoning Paths	67.8
+ Grounding	70.1
+ Brevity & Format Control	71.5
Omni-Zero-Doc (Full SIEGE)	**72.5**

Table 3. Ablation results for the SIEGE data engine.

5. Conclusion

Our work demonstrates that VLMs can self-evolve by generating their own visual data and questions. Our results across multiple model sizes show that stronger base models benefit more from self-evolution, as they produce higher-quality visual code from the start—the 7B and 8B models consistently outperform the 4B model. However, the primary limitation of this work is that we were unable to validate this scaling trend on larger VLMs (e.g., 38B parameters) due to the prohibitive computational cost, leaving the full scaling behavior of self-evolving training as an important direction for future work.

In conclusion, Omni-Zero-Doc establishes a transformative and highly scalable path for the zero-data self-evolution of Visual Language Models (VLMs), effectively dismantling the heavy reliance on human-annotated datasets that has long bottlenecked the field. By leveraging programmatic generation to synthesize an exhaustive array of complex visual layouts—ranging from intricate nested tables to non-linear academic formats—the framework ensures that models are exposed to a diversity of structural edge cases far beyond the scope of traditional datasets. This synthetic diversity is refined through a rigorous axiomatic filtering process, which serves as a logical gatekeeper to ensure that only high-fidelity, structurally coherent data facilitates the learning process, thereby bridging the critical structural reasoning gap in document understanding. Ultimately, Omni-Zero-Doc shifts the paradigm from passive data consumption to an active, self-improving loop of structural discovery, providing a robust blueprint for the next generation of autonomous document intelligence that scales with computational power rather than manual labor.

References

Clune, J. Ai-gas: Ai-generating algorithms, an alternate paradigm for producing general artificial intelligence. *arXiv preprint arXiv:1905.10985*, 2019.

Fei, H., Yao, Y., Zhang, Z., Liu, F., Zhang, A., and Chua, T.-S. From multimodal llm to human-level ai: Modal-

ity, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pp. 1–8, 2024.

Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., and Yue, X. Video-r1: Reinforcing video reasoning in mllms, 2025. URL https://arxiv.org/abs/2503.21776.

He, Y., Huang, C., Li, Z., Huang, J., and Yang, Y. Visplay: Self-evolving vision-language models from images. *arXiv preprint arXiv:2511.15661*, 2025.

Huang, C., Yu, W., Wang, X., Zhang, H., Li, Z., Li, R., Huang, J., Mi, H., and Yu, D. R-zero: Self-evolving reasoning llm from zero data. 2025a. URL https://arxiv.org/abs/2508.05004.

Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025b.

Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., and Shi, G. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1578–1597, 2025. doi: 10.1109/CVPRW67362.2025.00147.

Liu, F., Guan, T., Li, Z., Chen, L., Yacoob, Y., Manocha, D., and Zhou, T. HallusionBench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv:2310.14566*, 2023a.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023b.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023c. URL https://arxiv.org/abs/2304.08485.

Schmidhuber, J. Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial general intelligence*, pp. 199–226. Springer, 2007.

Thawakar, O., Venkatraman, S., Thawkar, R., Shaker, A., Cholakkal, H., Anwer, R. M., Khan, S., and Khan, F. Evolmm: Self-evolving large multimodal models with continuous rewards. *arXiv preprint arXiv:2511.16672*, 2025.

Wang, H., Yang, Y., Hu, J., Zhu, M., and Chen, W. V-zero: Self-improving multimodal reasoning with zero annotation. *arXiv preprint arXiv:2601.10094*, 2026.

Xiang, Z., Yang, C., Chen, Z., Wei, Z., Tang, Y., Teng, Z., Peng, Z., Li, Z., Huang, C., He, Y., et al. A systematic survey of self-evolving agents: From model-centric to environment-driven co-evolution.

Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., and Yuan, L. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

Yue, Z., Upasani, K., Yang, X., Ge, S., Nie, S., Mao, Y., Liu, Z., and Wang, D. Dr. zero: Self-evolving search agents without training data, 2026. URL https://arxiv.org/abs/2601.07055.