AIDE: AGENTICALLY IMPROVE VISUAL LANGUAGE MODEL WITH DOMAIN EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The enhancement of Visual Language Models (VLMs) has traditionally relied on knowledge distillation from larger, more capable models. This dependence creates a fundamental bottleneck for improving state-of-the-art systems, particularly when no superior models exist. We introduce AIDE (Agentic Improvement through Domain Experts), a novel framework that enables VLMs to autonomously enhance their capabilities by leveraging specialized domain expert models. AIDE operates through a four-stage process: (1) identifying instances for refinement, (2) engaging domain experts for targeted analysis, (3) synthesizing expert outputs with existing data, and (4) integrating enhanced instances into the training pipeline. Experiments on multiple benchmarks, including MMMU, MME, MMBench, etc., demonstrate AIDE's ability to achieve notable performance gains without relying on larger VLMs. Our framework provides a scalable, resource-efficient approach to continuous VLM improvement, addressing critical limitations in current methodologies, particularly valuable when larger models are unavailable or impractical to access.

1 INTRODUCTION

Visual Language Models (VLMs) have achieved impressive advancements in understanding and reasoning about visual content (Jean-Baptiste et al., 2022; Haotian et al., 2023; Fang et al., 2024). However, their continued improvement often hinges on knowledge distillation from larger, more capable models through approaches like instruction tuning(Fuxiao et al., 2023; Haotian et al., 2023). While this approach has proven effective for intermediate-scale models, it introduces a significant limitation for the largest state-of-the-art systems: the absence of a superior model renders further enhancement infeasible. This "chicken-and-egg" problem stifles progress and raises a critical question: how can VLMs be improved when no superior models exist?

Despite their general capabilities, VLMs frequently underperform in specialized tasks compared to domain expert models such as object segmentation tools or Optical Character Recognition (OCR) systems. For instance, models like Grounding DINO(Shilong et al., 2023) consistently outperform general-purpose VLMs(Lu et al., 2021; Shaohan et al., 2023) in visual recognition tasks (Table 1). This observation suggests an alternative pathway: rather than relying on larger general models, VLMs can leverage the specialized capabilities of expert models for improvement.

In this paper, we introduce AIDE (Agentic Improvement through Domain Experts), a framework that enables VLMs to strategically collaborate with domain expert models to enhance their training data. AIDE employs a four-stage workflow: (1) identifying instances requiring refinement, (2) invoking expert models for specialized outputs, (3) synthesizing these outputs with existing data, and (4) systematically integrating improved data points into the training process.

We validate AIDE's effectiveness through extensive experiments on benchmarks such as MMMU(Yue et al., 2024), MME(Fu et al., 2024), MMBench(Liu et al., 2024), etc., showing that it achieves notable performance improvements using only off-the-shelf lightweight expert models. Unlike traditional methods, AIDE does not depend on access to larger models, making it a scalable and computationally efficient solution for advancing state-of-the-art VLMs.

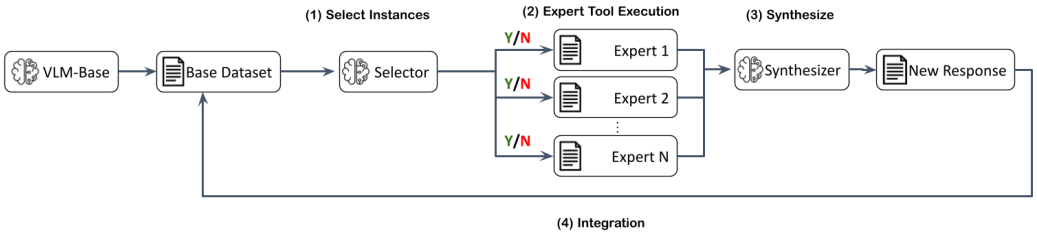

Figure 1: AIDE workflow. The framework consists of two agents, a Selector and a Synthesizer. The Selector interacts with the data instances and autonomously invoke the expert tools as it thinks fit. The Synthesizer collects all information from the original data instances along with outputs from the select experts, and then generate an enriched response.

2 Related Work

Knowledge Distillation and Self-Improvement Traditional methods for improving VLMs rely on knowledge distillation, where a larger "teacher" model generates training data to enhance the performance of a smaller "student" model. While effective for intermediate-scale models, this paradigm creates a dependency on the availability of superior models, which limits its applicability to state-of-the-art systems. Approaches like recursive self-improvement(**?**) have shown promise in leveraging larger models for training smaller ones, but their scalability may be inherently constrained.

Specialized Models and Multimodal Data Recent studies highlight the superiority of domain-specific expert models in certain tasks. For example, object detection systems such as Grounding DINO and OCR models like PaddleOCR significantly outperform general-purpose VLMs in their respective domains. These findings underscore the potential of leveraging specialized models to complement the general capabilities of VLMs.

Data Synthesis and Augmentation Existing methods for augmenting training data often involve the model itself generating synthetic examples. While this approach can enhance performance on specific benchmarks, it risks perpetuating the biases and limitations of the model, resulting in diminishing returns. By contrast, AIDE integrates external expert knowledge into the training pipeline, enabling more robust and unbiased improvements.

Table 1: Performance comparison between general-purpose models and expert models on referring expression comprehension tasks.

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
general-purpose model								
Kosmos-2	52.3	57.4	47.3	45.5	50.7	42.2	60.6	61.7
Florence-2-B	53.9	58.4	49.7	51.5	56.4	47.9	66.3	65.1
Florence-2-L	56.3	61.6	51.4	53.6	57.9	49.9	68.0	67.0
specialist model								
Grounding DINO L	**90.56**	**93.19**	**88.24**	**82.75**	**88.95**	**75.92**	**86.13**	**87.02**

3 AIDE Framework

The AIDE framework enables VLMs to autonomously improve by collaborating with domain expert models. It comprises two primary agents—Selector and Synthesizer—and operates through three principal actions: Selection, Execution, and Synthesis. Fig. 1 provides an overview of the AIDE pipeline.

Selector The selector serves two objectives, identify improvement candidates and match candidates with expert tools: the selector interacts with the base dataset and is presented with detailed information and functionalities of the expert tools and judge if any of the additional information the experts can provide may be beneficial to improve the quality of the data. If it is then the selector will exercise the corresponding expert tool.

Synthesizer The Synthesizer integrates expert outputs with the original data to generate enhanced training examples. This process involves:

- Aggregating information from multiple sources.
- Resolving potential conflicts.

Table 2: AIDE performance improvements. Using VLM for self improvement enhanced the performance on various benchmarks.

Selector	chartqa	textvqa	mathvista	mmbench	mme-c	mme-p	mmmu_val	sciQA_img	pope
Baseline (reproduced)	0.8024	0.7469	55.8	74.055	337.5	1529.0908	0.41	0.8042	0.8929
Hardcode5	0.7964	0.4149	54.5	71.0481	336.4286	1511.9161	0.409	0.7705	0.8956
Eagle (single)	0.7964	0.7685	56.6	74.3986	385.7143	1590.0183	0.414	0.8299	0.8919
Eagle (double)	0.804	0.7446	56.9	74.8282	335.7143	1583.5986	0.422	0.7883	0.9007

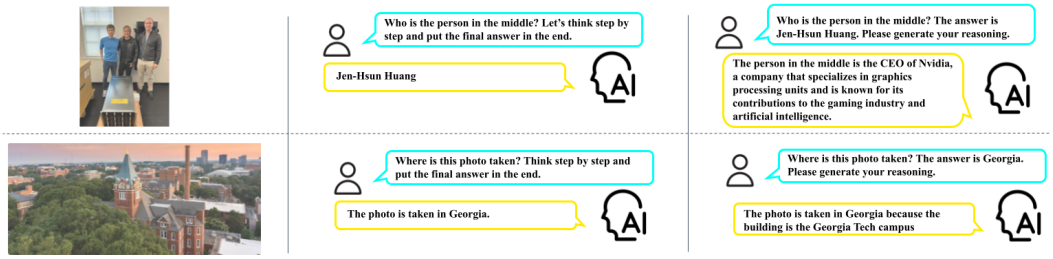

Figure 2: Small-step prompting. We observe even when VLM is able to answer the query (*middle-column*), sometimes the instruction following is not stable. And simplifying the prompt into smaller steps by giving the answer (*last column*) gives better results.

- Producing richer and more coherent responses.

3.1 Integration Workflow

After generating enhanced examples, the Update phase incorporates them into the training pipeline. This involves:

Filtering : Ensures new formulations maintain sensible information with original instances.

Ground Truth Retention: Preserving original ground truth data to prevent catastrophic forgetting.

4 Experiment

Setup We evaluate AIDE using the Eagle-8B(Shi et al., 2024) model as both Selector and Synthesizer, interacting with the Cambrian1-7M dataset. Experiments are conducted on an NVIDIA A100 node with 8 GPUs. Note that the choice of Selector and Synthesizer cam be adaptable and need not be the same.

Expert Tool Choice Two lightweight domain experts, PaddleOCR() and Grounded-SAM(Tianhe et al., 2024), are employed. These tools complement the visual data-rich composition of Cambrian1-7M(Shengbang et al., 2024). AIDE is extensible to incorporate additional expert models for multi-modal tasks.

Results Table 2 shows that applying AIDE is able to improve on MMMU by 1.2%, MMBench by 0.77%, MME by 52, Mathvista by 1.1 %, etc. These results highlight AIDE's effectiveness in leveraging domain expertise for VLM improvement.

4.1 Ablations

We evaluated variations in Selector strategies, including text-only LLMs and heuristic methods. While these methods yielded modest improvements, their impact was limited to specific data types, such as charts. Retaining ground truth data during updates was critical in preventing catastrophic forgetting.

multi-step prompting In the synthesize step, we tried to directly prompt the VLM to do leverage the expert information and put final answer at the end. But often failed to do so. Fig. 2 show that it is better to have multi-step prompting agentic workflow instead of directly prompt the CoT. If we directly ask CoT, doesn't work well. But if we simplify to just generate the reasoning, and then append the ground truth to ensure faithfulness.

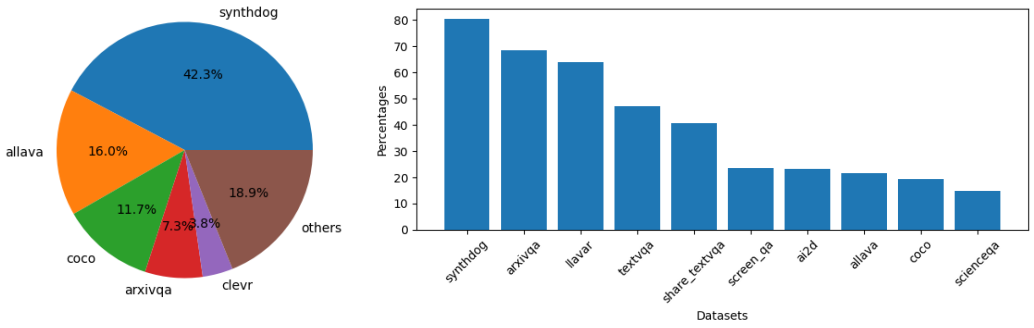

Figure 3: *Left:* Breakdown of selected data instances by VLM-Selector. Synthdog takes the most proportion of the selection. *Right:* Percentage by selected data instances by VLM-Selector to the multimodal source from original Cambrian-1. 80% of the synthdog is selected by AIDE framework.

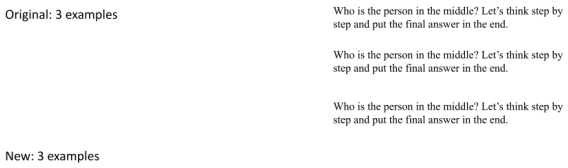

Figure 4: Comparison of relabeled dataset. AIDE workflow makes the dataset better at reasoning.[Ming: will find some examples later]

4.2 Analysis on AIDE-selected data

We analyze the VLM selected data points for improvement. Out of the 7M training instance, about 2M are text-only and 5M are multimodal. And of the multimodal training instances, around 950K were selected by VLM-Selector. We provide the breakdown of the 950K in Tab. **??**. Detailed analysis of AIDE-selected data reveals several interesting patterns. In two directions, the proportions chosen among the 950k and proportions among the multimodal samples in Cambrian1

Synthdog is selected for improvement mostly We can see that the majority (over 40%) of the selected are from synthdog, an OCR dataset, meaning VLM thinks the quality of synthdog is bad.

Synthdog are chosen but some document data are also selected On the other hand, we analyze the percentage of the selected candidates compare to the original data from Cambrian1. Again, most of the synthdog are chosen by the selector for improvement, arxivqa llavar, textvqa are also predominately selcted, meaning...

4.3 Qualitative results

Figure 4 illustrates the enriched data instances compared to the original data. These enhancements corroborate AIDE's ability to address shortcomings and improve contextual understanding.

5 Conclusion

We presented AIDE, an agentic framework enabling VLM improvement through domain expert models. Unlike traditional methods, AIDE offers a scalable, resource-efficient alternative to reliance on larger models. Our contributions include:

- A novel approach to VLM enhancement without superior models.
- Demonstrated improvements across benchmarks like MMMU, MMBench, and SciQA.
- Detailed analysis of data selection strategies and their impacts.

Future work may explore adapting AIDE for preference optimization, generating new (question, answer) pairs and incorporating test-time inference techniques to further enhance the quality of new data. These advancements aim to further refine synthesized data quality and broaden AIDE's applicability, paving the way for continuous VLM training paradigms.

REFERENCES

Yunhao Fang, Ligeng Zhu, Yao Lu, Yan Wang, Pavlo Molchanov, Jan Kautz, Jang Hyun Cho, Marco Pavone, Song Han, and Hongxu Yin. Vila^2: Vila augmented vila. *arXiv preprint arXiv:2407.17453*, 2024. URL https://www.arxiv.org/abs/2407.17453.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. URL https://arxiv.org/abs/2306.13394.

Liu Fuxiao, Lin Kevin, Li Linjie, Wang Jianfeng, Yacoob Yaser, and Wang Lijuan. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. URL https://www.arxiv.org/abs/2306.14565.

Liu Haotian, Li Chunyuan, Wu Qingyang, and Lee Yong, Jae. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. URL https://www.arxiv.org/abs/2304.08485.

Alayrac Jean-Baptiste, Donahue Jeff, Luc Pauline, Miech Antoine, Barr Iain, Hasson Yana, Lenc Karel, Mensch Arthur, Millican Katie, Reynolds Malcolm, Ring Roman, Rutherford Eliza, Cabi Serkan, Han Tengda, Gong Zhitao, Samangooei Sina, Monteiro Marianne, Menick Jacob, Borgeaud Sebastian, Brock Andrew, Nematzadeh Aida, Sharifzadeh Sahand, Binkowski Mikolaj, Barreira Ricardo, Vinyals Oriol, Zisserman Andrew, and Simonyan Karen. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. URL https://www.arxiv.org/abs/2204.14198.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Yuan Lu, Chen Dongdong, Chen Yi-Ling, Codella Noel, Dai Xiyang, Gao Jianfeng, Hu Houdong, Huang Xuedong, Li Boxin, Li Chunyuan, Liu Ce, Liu Mengchen, Liu Zicheng, Lu Yumao, Shi Yu, Wang Lijuan, Wang Jianfeng, Xiao Bin, Xiao Zhen, Yang Jianwei, Zeng Michael, Zhou Luowei, and Zhang Pengchuan. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. URL https://www.arxiv.org/abs/2111.11432.

Huang Shaohan, Dong Li, Wang Wenhui, Hao Yaru, Singhal Saksham, Ma Shuming, Lv Tengchao, Cui Lei, Mohammed Owais, Khan, Patra Barun, Liu Qiang, Aggarwal Kriti, Chi Zewen, Bjorck Johan, Chaudhary Vishrav, Som Subhojit, Song Xia, and Wei Furu. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. URL https://www.arxiv.org/abs/2302.14045.

Tong Shengbang, Brown Ellis, Wu Penghao, Woo Sanghyun, Middepogu Manoj, Akula Sai, Charitha, Yang Jihan, Yang Shusheng, Iyer Adithya, Pan Xichen, Wang Ziteng, Fergus Rob, LeCun Yann, and Xie Saining. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860v2*, 2024. URL https://www.arxiv.org/abs/2406.16860v2.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv:2408.15998*, 2024.

Liu Shilong, Zeng Zhaoyang, Ren Tianhe, Li Feng, Zhang Hao, Yang Jie, Li Chunyuan, Yang Jianwei, Su Hang, Zhu Jun, and Zhang Lei. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. URL https://www.arxiv.org/abs/2303.05499.

Ren Tianhe, Liu Shilong, Zeng Ailing, Lin Jing, Li Kunchang, Cao He, Chen Jiayu, Huang Xinyu, Chen Yukang, Yan Feng, Zeng Zhaoyang, Zhang Hao, Li Feng, Yang Jie, Li Hongyang, Jiang Qing, and Zhang Lei. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. URL https://www.arxiv.org/abs/2401.14159.

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*, pp. 9556–9567, 2024.