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 (Alayrac et al., 2022; Liu et al., 2023b; Fang et al., 2024). However, their continued improvement often hinges on knowledge distillation from larger, more capable models through approaches like instruction tuning (Liu et al., 2023a;b). 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 often 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 (Liu et al., 2023c) consistently outperform general-purpose VLMs (Yuan et al., 2021; Huang 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 (Xiang et al., 2024), MME (Fu et al., 2024), MMBench (Yuan 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.

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

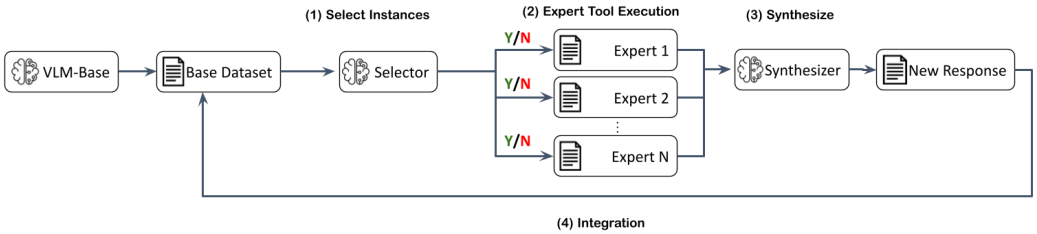

Figure 1: AIDE workflow. The framework consists of two agents, a Selector and a Synthesizer. The Selector interacts with the data instances and autonomously invoke the expert tools as it deems fit. The Synthesizer collects all information from the original data instances along with outputs from the select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

performance of a smaller "student" model. While effective for intermediate-scale models (Liu et al., 2023b;a), this paradigm creates a dependency on the availability of superior models, which limits its applicability to state-of-the-art systems.

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 (Yuan et al., 2021; Ren et al., 2024; Liu et al., 2023c). 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 generating synthetic examples (Liu et al., 2023a). While this approach can enhance performance on specific benchmarks, it risks perpetuating the biases and limitations of the model, resulting in diminishing returns. In contrast, AIDE integrates external expert knowledge and the original samples into the data generation 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. AIDEpresumes an existing base dataset as the environment for agents to interact with. 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, i.e. the original instances and domain expert outputs.
- Resolving potential conflicts. E.g. between original instances and expert outputs.
- Producing richer and more coherent responses.

By these instructions, we expect the new response s would inherently be richer and contain reasoning flavors. See Sec. 4.3 for comparisons.

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
Heuristic (ans $\leq$5 tokens)	0.7964	0.4149	54.5	71.0481	336.4286	1511.9161	0.409	0.7705	0.8956
Eagle-single (Ours)	0.7964	**0.7685**	56.6	74.3986	**385.7143**	**1590.0183**	0.414	**0.8299**	0.8919
Eagle-double (Ours))	**0.804**	0.7446	**56.9**	**74.8282**	335.7143	1583.5986	**0.422**	0.7883	**0.9007**

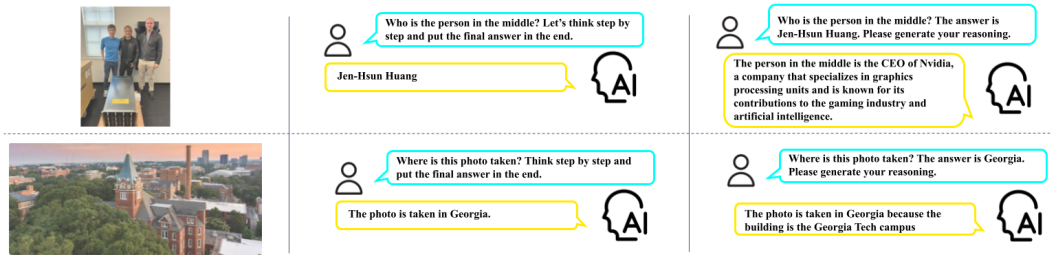

Figure 2: Small-step prompting. We observe even when VLM is able to answer the query (*middle-column*), sometimes the instruction following is not stable. And simplifying the prompt into smaller steps by giving the answer (*last column*) gives better results.

3.1 INTEGRATION

After generating enhanced samples, the integration incorporates them back into the training pipeline. This involves filtering: ensures new formulations maintain sensible information along with the original instances to prevent model collapse.

4 EXPERIMENT

Setup We evaluate AIDE using the Eagle-8B (Shi et al., 2024) as both Selector and Synthesizer, interacting with the Cambrian1-7M dataset for one iteration. 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 (pad) and Grounded-SAM (Ren et al., 2024), are employed. These tools complement the visual data-rich composition of Cambrian1-7M (Tong et al., 2024). AIDE is extensible to incorporate additional expert models for multimodal tasks.

Integration We use simple heuristics like n-gram filtering, etc. because we use small-step prompting that we deem enough to maintain quality of new responses (Fig. 2 & Fig. 4), but AIDE can easily add verifiers to further enhance data generation quality (see Sec. 5 for discussions).

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

Selector choice We evaluated variations in Selector strategies, including text-only LLMs and heuristic methods. Tab. 2 (row 2) shows that heuristic like synthesizing the instances that has $\leq$ 5 tokens is not comparable to using a VLM-selector, even though the heuristic would select much more instances for synthesis (2.5M vs 950k)

Small-step prompting In the synthesis step, we tried to directly prompt the VLM to generate more detailed responses and then put final answer at the end, but it often fails to do so (Fig. 2-*mid*). Fig. 2-*last* shows that it is effective to simply prompt the VLM with the whole information of the instance and with one task (e.g. just generate the reasoning). Even though the VLM knows the correct answers, slightly more complex prompt cannot achieve the desirable outcome.

Originals retention Tab. 2 shows that AIDE is able to provide improvements with or without the original turn (question, answer pair), dubbed *single* and *double* respectively, suggesting the effectiveness of AIDE.

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

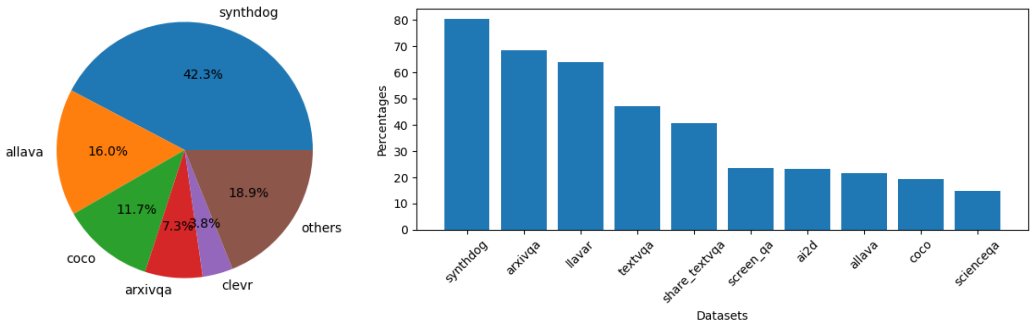

Figure 3: *Left:* Breakdown of selected data instances by VLM-Selector. Synthdog takes the most proportion of the selection. *Right:* Ratio of data instances selected by the VLM-Selector to the total instances in the original Cambrian-1.

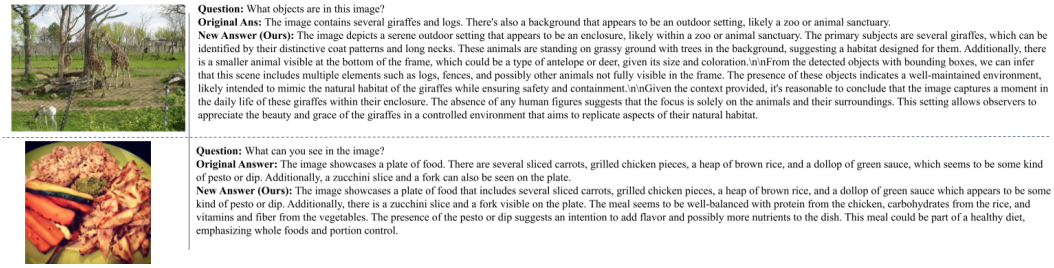

Figure 4: Comparisons of the original and the new answer produced by AIDE . Our AIDE workflow enriches the responses.

were selected by VLM-Selector. We provide the breakdown of the 950K in Fig. 3. 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 takes the most portion among selected instances. We observe that the majority (over 40%) of the selected are from synthdog, an OCR dataset, suggesting the Selector deems the quality of synthdog need most improvement.

Most synthdog are chosen and other document data are also selected On the other hand, we analyze the percentage of the selected candidates compared to the original data by source from Cambrian1. Again, about 80% of the synthdog are chosen by the VLM selector for improvement. Arxivqa llavar, textvqa are also predominately selected. We suspect the Selector deems the quality of these document datasets need improvement and posit AIDE may serve as an alternative way to estimate the quality of a dataset through a VLM-as-a-judge approach.

4.3 QUALITATIVE RESULTS

Figure 4 illustrates the comparisons between the original data instances and the enriched data instances by our AIDE workflow. The new responses provides more details and reasoning-flavored context than the original answers. These enhancements may explain AIDE's ability to improve the performances on various benchmarks (Tab. 2).

5 DISCUSSION

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 quarantee 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

PaddleOCR. https://github.com/PaddlePaddle/PaddleOCR. Accessed: 2024-06-30.

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

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.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 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.

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

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

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

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

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.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. 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.

Yue Xiang, Ni Yuansheng, Zhang Kai, Zheng Tianyu, Liu Ruoqi, Zhang Ge, Stevens Samuel, Jiang Dongfu, Ren Weiming, Sun Yuxuan, 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.

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

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