LATTE: Learning to Reason with Vision Specialists

Anonymous CVPR submission

Paper ID 36

Abstract

While open-source vision-language models perform well on simple question-answering, they still struggle with complex questions that require heterogeneous vision capabilities. Unfortunately, we have yet to develop methods that infuse fine-grained recognition, visual grounding, depth estimation, and 3D reasoning into a single vision-language model. Instead of forcing smaller models to learn both perception and reasoning, we propose LATTE*, a family of vision-language models that have LeArned to Think wiTh vision spEcialists. By offloading perception to state-of-the-art vision models, our approach enables vision-language models to focus solely on reasoning over high-quality perceptual information. To train* LATTE*, we create and filter a large dataset of 273K high-quality synthetic reasoning traces over perceptual outputs of vision specialists.* LATTE *trains on this data and brings significant gains across 6 benchmarks covering both perception and reasoning abilities, compared to baselines instruction-tuned with direct answers. On the other hand, models trained by distilling both perception and reasoning from larger models lead to smaller gains or even degradation on some perception tasks. Further, our method results in a* 2% *to* 5% *improvement on average across all benchmarks over the vanilla instruction-tuned baseline regardless of model backbones, with gains up to* 16% *in MMVet.*

1. Introduction

The landscape of real-world vision-language tasks is vast, spanning from basic visual question answering [1] and fine-grained object recognition to complex multi-step geometric reasoning [8]. These tasks demand both perception and reasoning. For instance, a user might photograph a gas price panel and ask how much fuel they can afford within a given budget (Figure 1). Solving this requires a vision-language model with strong perception—localizing prices via OCR—and multi-step reasoning to compute the answer. While large proprietary models like GPT-4o excel due to extensive data and model size scaling, smaller open-source models still struggle [22].

To narrow the gap between large proprietary models and smaller open-source counterparts within a reasonable budget, researchers have explored distilling both perception and reasoning from larger vision-language models [25, 28] or specialized vision models [9]. Despite these efforts, open-source models continue to lag behind.

We argue that the primary reason for this lag is the perception limitations of open-source vision-language models. While open-source language models have largely caught up with their proprietary counterparts [2, 13], vision remains a complex fusion of heterogeneous capabilities. The computer vision community has historically tackled these capabilities separately—e.g., DepthAnything [29] for depth estimation and GroundingDINO [19] for object recognition—while unified models still lag behind [20]. Similarly, the human brain dedicates distinct regions to categorical recognition (ventral stream) and spatial reasoning (dorsal stream)[6], with the reasoning and language-processing frontal and temporal lobes occupying a smaller volume[12]. By contrast, vision-language models remain heavily skewed toward language, treating visual encoders as an afterthought [4].

We depart from the *learning to perceive and reason* paradigm to propose a new approach: *learning to reason with vision specialists*. Rather than expecting a small model to master both perception and reasoning, we leverage decades of advancements in computer vision by relying on specialized vision models to provide perceptual information. This allows the vision-language model to focus exclusively on acquiring perceptual information from vision specialists and reasoning over them—enabling it to '*see further by standing on the shoulders of giants*.' Such a paradigm reduces the burden on models to extract low-level perceptual signals, allowing them to concentrate on higher-level reasoning while benefiting from the robust capabilities of dedicated vision specialists, which is particularly important for small open-source models because of their limited capacity to effectively learn both perception and reasoning.

To implement this paradigm, we curate high-quality training data in the form of multi-step reasoning traces that integrate perceptual information from vision specialists. We

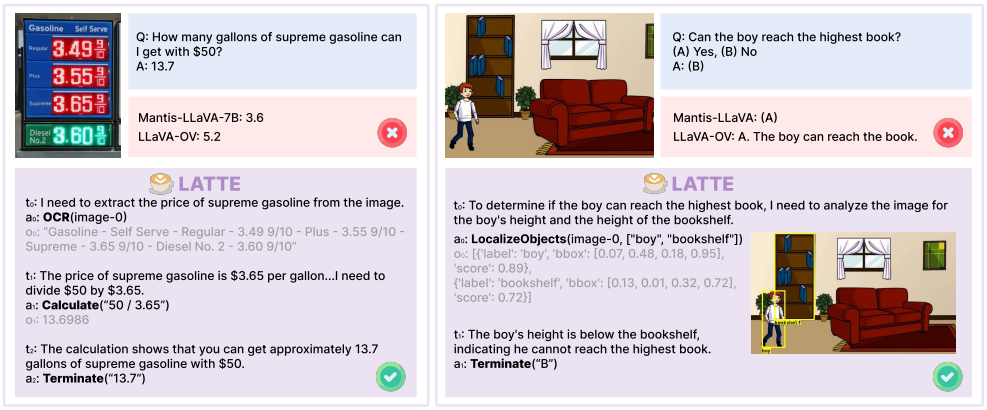

Figure 1. **Example outputs of LATTE vs. SoTA multi-modal large language models.** Our LATTE model is able to answer challenging visual questions by reasoning over perceptual information output by vision specialists.

formulate the multi-step reasoning traces as LATTE-trace, where each step consists of: (1) a *thought* for verbalized reasoning; (2) an *action* to retrieve perceptual information from a specific vision specialist; and (3) an *observation* of the returned data. Since obtaining these traces at scale with human annotators is costly, we develop two data engines for synthetic data generation. First, we leverage GPT-4o's strong multimodal reasoning and state-of-the-art vision specialists' precise perception to generate large-scale synthetic reasoning traces across diverse image sources, applying aggressive filtering and mixing techniques. Second, we generate reasoning traces using Python programs and structured reasoning templates, comparing them against GPT-generated traces to evaluate reasoning quality. In total, we produce over 1M reasoning traces across 31 datasets with GPT-4o and handcrafted programs.

With this data, we finetune small multi-modal language models to reason with vision specialists and evaluate our models on 6 benchmarks covering both perception and reasoning skills. We compare our model to two types of baselines: (1) multi-modal language models trained with vanilla instruction tuning with only direct answers; and (2) models trained by distilling both perception and reasoning.

Finally, we highlight four major takeaways from our experiments: First, learning to reason with vision specialists enables our model to outperform vanilla instruction-tuned baseline by significant margins on both perception and reasoning benchmarks, with an overall average gain of 6.4%. By contrast, the other distillation methods lead to smaller gains or even degradation in the perception performance. This trend holds as we scale the training data. Second, our method consistently outperforms the vanilla instruction-tuned baseline by $2-5\%$ on average across all benchmarks regardless of model backbones, with staggering performance gains of $10-20\%$ on MMVet. Third, through data ablations, we confirm that the quality of LATTE-trace matters more than quantity: our best data recipe consists of only 293K LATTE-trace which GPT-4o generated and answered correctly, and it leads to larger performance gains than all other data recipes of larger scales (up to 2x larger or more). Finally, programmatically-generated LATTE-trace can hurt model performance as a result of the worse reasoning quality, suggesting that again that high-quality reasoning is crucial to the model's performance.

2. Learning to Think with Vision Specialists

Our goal is to train vision-language models to reason about complex multi-modal tasks with the help of vision specialists. To train such models, we need reasoning traces that involve (1) invoking vision specialists and (2) reasoning over their outputs. We refer to such data as LATTE-trace. One LATTE-trace $\mathcal{T}$ is a sequence of steps S_i, where each step consists of thought t_i, action a_i and observation o_i:

$$\mathcal{T} = (S_0, S_1, ..., S_n) = (S_i)_{i=0}^{n} \quad (1)$$

$$S_i = (t_i, a_i, o_i), t_i \in L, a_i \in A \quad (2)$$

where L represents language space, and A is the action space consisting of vision specialists. Note that the model only generates t_i and a_i, which the training loss is applied on, whereas o_i is obtained from the vision specialists.

Action space. The action space A of our model consists of vision tools that are either specialized vision models or image processing tools. Concretely, these include OCR [10], GETOBJECTS [33], LOCALIZEOBJECTS [19], ESTIMATEOBJECTDEPTH [29], ESTIMATEREGIONDEPTH [29], DETECTFACES [17], CROP, ZOOMIN, GETIMAGETOTEXTSSIMILARITY [23], GETIMAGETOIMAGESSIMILARITY [23], GETTEXTTOIMAGESSIMILARITY [23]. Inspired by prior works on multi-modal tool use [7, 8, 18, 22, 26], we include a few additional tools to help with reasoning: QUERYLANGUAGEMODEL, QUERYKNOWLEDGEBASE, CALCULATE, and SOLVEMATHEQUATION. We also include TERMINATE as a tool for the model to output a final answer in the same action format. Our final action space consists of 15 tools, and their full implementation details can be found in the Appendix.

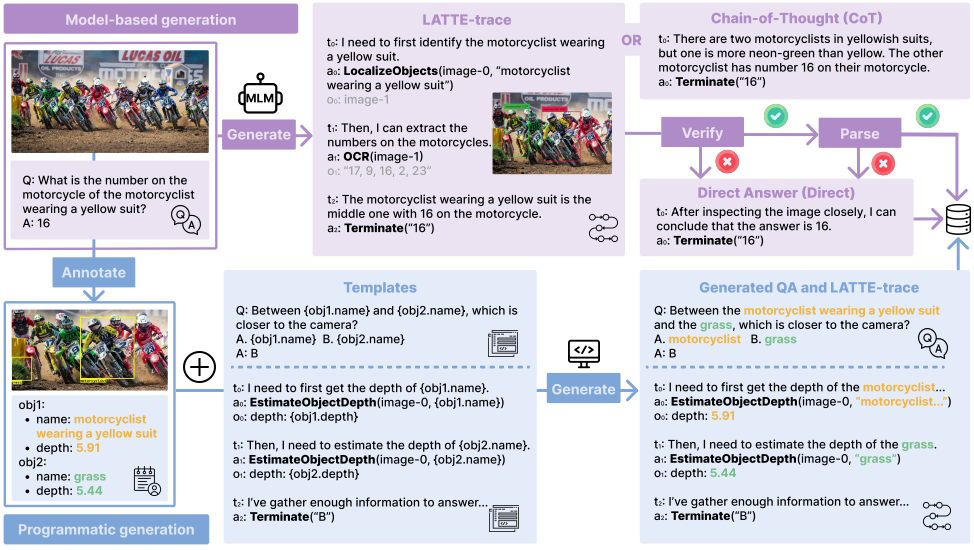

Figure 2. **Data generation.** We illustrate our model-based data generation (top) and programmatic generation (bottom) pipelines.

2.1. LATTE-trace generation

We generate synthetic LATTE-trace data with two automatic approaches: Model-based generation and Programmatic data generation.

Model-based generation. The model-based data generation pipeline consists of three steps (Figure 2 top):

1. GENERATE. First, we leverage images and QA examples in existing visual instruction tuning datasets and generate LATTE-traces to solve the questions with GPT-4o (2024-08-06). We include diverse questions on both single-image and multi-image examples from two large-scale instruction tuning datasets, Cauldron and Mantis-Instruct [11, 14]. We feed the images and questions to GPT-4o and prompt it to answer the questions by following a LATTE-trace or just CoT when it is not necessary (*e.g.*, the question is straightforward) or helpful (*e.g.*, the question requires domain-specific knowledge out of the scope of available tools) to call specialized vision tools (Figure 2).
2. VERIFY. Second, we verify GPT-4o's generated answers against the ground-truth. We force GPT-4o to always end with TERMINATE(answer) and compare its prediction to the ground-truth. If the final answer following a reasoning trace is correct, we move this LATTE-trace to the next stage. Otherwise, we convert this example into the direct answer (Direct) format with the ground-truth (Figure 2).
3. PARSE. Finally, we check the JSON syntax of each step of the LATTE-trace. Similar to the previous stage, we again keep the LATTE-traces free of syntax errors and turn the others into the Direct format with ground-truth answers.

Programmatic data generation. While model-based data generation distills reasoning from proprietary models, we are curious if reasoning with vision specialists can be learned in another manner without reliance on proprietary models. To study this perspective, we implement a programmatic data generation engine for synthesizing LATTE-traces (Figure 2 bottom). This pipeline involves two steps:

1. ANNOTATE. First, we gather existing dense annotations of images. We adopt Visual Genome (VG) as it contains rich human annotations of objects, attributes, and relationships of the images. In addition, we obtain depth maps of the VG images with Depth-Anything-v2 [29].
2. GENERATE. Next, we programmatically generate both the QA pairs and the corresponding LATTE-traces with manually written templates and the dense annotations of the images. We reuse the pipeline from [31, 32] for generating diverse QA pairs that cover various vision capabilities such as counting and spatial understanding. To generate LATTE-traces, we define templates for thoughts, actions, and observations across all steps. See Appendix for more details.

3. Experiments

We perform extensive experiments with small multi-modal models and 9 data recipes on 6 benchmarks.

Models. We adopt models that support multi-image inputs as our data includes reasoning traces with multiple images. For most of our experiments, we use Mantis-8B-SigLIP-LLaMA-3 as the base model. We additionally experiment with Mantis-8B-CLIP-LLaMA-3, and LLaVA-OneVision-7B (Qwen2-7B and SigLIP) in our ablations.

Baselines. We compare our model to two types of baselines: (1) vanilla instruction-tuning (Vanilla IT) – instruction-tuning data with only direct answers – and (2) distillation methods that train small models by distilling both perception and reasoning from larger models, including VPD [9], VisCoT [25], and LLaVa-CoT[28].

Evaluation setup. We select 6 multi-modal benchmarks covering both perception and reasoning. The perception-focused benchmarks include RealWorldQA, CV-Bench and

Table 1. **LATTE vs. Baselines on Perception and Reasoning Benchmarks.** Our method LATTE brings substantial gains over the vanilla instruction-tuned (Vanilla IT) baseline on both perception and perception + reasoning benchmarks.

Method	Perception				Perception + Reasoning				Overall
	BLINK	CV-Bench	RealWorldQA	Avg	MathVista	MMStar	MMVet	Avg	Avg
Vanilla IT	44.1	49.2	41.4	44.9	31.0	39.7	27.8	32.8	38.9
VPD	41.6	48.8	**44.8**	45.1 (+0.2)	33.0	41.1	32.8	35.7 (+2.8)	40.4 (+1.5)
LLaVa-CoT	42.2	40.4	38.0	40.2 (-4.7)	36.7	**44.6**	40.2	40.5 (+7.7)	40.4 (+1.5)
LATTE	**46.4**	**54.0**	42.0	**47.5** (+2.6)	**36.9**	44.2	**47.9**	**43.0** (+10.2)	**45.2** (+6.4)

Table 2. **LATTE vs. Vanilla IT with Different Models.** We learn that LATTE leads to performance gains over Vanilla IT regardless of the base models. The gains are 2-5% on average across all 6 benchmarks and up to 16% on MMVet.

Language / Vision	Starting checkpoint	Method	Perception				Perception + Reasoning				Overall
			CV-Bench	BLINK	RealWorldQA	Avg	MathVista	MMStar	MMVet	Avg	Avg
LLaMA3-8B / CLIP	Mantis Pretrained	Vanilla IT	52.6	45.8	52.3	50.2	33.1	36.7	28.9	32.9	41.6
		LATTE	56.9	49.6	51.1	52.6	36.6	40.8	45.2	40.8	46.7 (+5.1)
LLaMA3-8B / SigLIP	Mantis Pretrained	Vanilla IT	52.3	43.7	51.8	49.3	31.1	40.5	33.0	34.9	42.1
		LATTE	57.2	47.8	53.7	52.9	34.9	44.6	45.2	41.6	47.2 (+5.2)
	Mantis Instruct-tuned	Vanilla IT	50.6	46.7	54.8	50.7	36.2	40.7	29.7	35.5	43.1
		LATTE	51.7	47.6	56.5	51.9	36.3	42.5	45.7	41.5	46.7 (+3.6)
Qwen2-7B / SigLIP	LLaVa-OV Stage 1.5	Vanilla IT	56.8	**50.3**	57.8	55.0	**42.4**	50.1	39.3	43.9	49.5
		LATTE	**60.2**	49.9	**58.8**	**56.3**	41.9	**51.0**	**50.9**	**48.0**	**52.1** (+2.7)

BLINK [5, 15, 24, 27], and the perception + reasoning ones are MathVista, MMStar, and MMVet [3, 21, 30]. Additional details can be found in the Appendix.

3.1. Main results

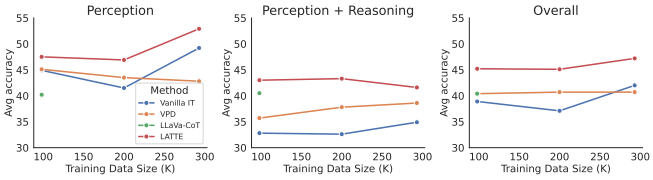

Figure 3. **Performance of LATTE vs. Baselines across Training Data Scales.** We find that our method leads to consistent gains across varying training data sizes – 98K, 200K and 293K.

Our method leads to substantial gains compared to vanilla instruction-tuning on both perception and reasoning benchmarks, whereas other distillation baselines result in smaller gains or even degradation on some perception tasks. We find that learning to reason with vision specialists enables our model to achieve consistent gains on perception-focused VQA benchmarks as well as benchmarks that require both perception and reasoning, with average gains of 2.6% and 10.2% respectively (Table 1). By contrast, both distillation baselines VPD and LLaVa-CoT bring much smaller gains, with an average of 1.5% across all benchmarks, compared to ours (6.4%). Further, we observe that the same trend holds as we scale the training data size from 98K to 200K and 293K, where our method consistently brings larger gains on both perception and perception + reasoning benchmarks (Figure 3). Interestingly, LLaVa-CoT even hurts the model's performance on perception benchmarks, even though it increases the performance on the perception + reasoning benchmarks (Table 1). This result suggests that GPT4-o might still be inferior to vision specialists on some perception tasks, as LLaVa-CoT distills purely from GPT4-o.

Our method beats the vanilla instruction-tuning baseline on average across all benchmarks regardless of the base model and checkpoint, with significant gains of 10-16% on MMVet. We fine-tune 3 different multi-modal models with all 293K LATTE-traces starting from different checkpoints. We observe that our method leads to consistent gains of 2-5% in the model's average accuracy across 6 benchmarks compared to the baselines instruction-tuned with the same examples in the Direct format (Table 2). We note that our method results in staggering gains of 10-16% on MMVet, which covers a wide range of perceptual and reasoning capabilities. Moreover, we find that our data results in larger gains on earlier pretrained checkpoints than on later-stage instruction-tuned checkpoints, likely due to the relatively small size of our data compared to Mantis' and LLaVa-OV's instruction-tuning data (1.2M and 4.5M) and some overlap in the images and questions [11, 16].

4. Conclusion

We propose to learn multi-modal language models to reason with vision specialists instead of becoming both vision specialists and reasoning experts.

Limitations and Future Work. First, our method requires customized implementations of the specialized vision tools. Second, reasoning with the vision specialists also requires additional compute at inference time. Future work can optimize and enhance the implementations of vision specialists.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[2] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. 1

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 4

[4] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1

[5] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 4

[6] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992. 1

[7] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. *ArXiv*, abs/2211.11559, 2022. 2

[8] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models, 2024. 1, 2

[9] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9601, 2024. 1, 3

[10] JadedAI. Easyocr, 2025. 2

[11] Dongfu Jiang, Xuan He, Huaye Zeng, Con Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 3, 4

[12] Georg B Keller, Tobias Bonhoeffer, and Mark Hübener. Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron*, 74(5):809–815, 2012. 1

[13] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. 1

[14] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 3

[15] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023. 4

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 4

[17] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: Dual shot face detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[18] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. Llava-plus: Learning to use tools for creating multimodal agents, 2023. 2

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2

[20] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024. 1

[21] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024. 4

[22] Zixian Ma, Weikai Huang, Jieyu Zhang, Tanmay Gupta, and Ranjay Krishna. m&m's: A benchmark to evaluate tool-use for multi-step multi-modal tasks. *EECV 2024*, 2024. 1, 2

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[24] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. 4

[25] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models, 2024. 1, 3

[26] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023. 2

[27] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. 4

[28] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2025. 1, 3

[29] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 1, 2, 3

[30] 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. In *International conference on machine learning*. PMLR, 2024. 4

[31] Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Task me anything. *arXiv preprint arXiv:2406.11775*, 2024. 3

[32] Jieyu Zhang, Le Xue, Linxin Song, Jun Wang, Weikai Huang, Manli Shu, An Yan, Zixian Ma, Juan Carlos Niebles, Silvio Savarese, Caiming Xiong, Zeyuan Chen, Ranjay Krishna, and Ran Xu. Provision: Programmatically scaling vision-centric instruction data for multimodal language models. *Preprint*, 2024. 3

[33] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 2