COMPACT: COMPositional Atomic-to-Complex Visual Capability Tuning

Anonymous CVPR submission

Paper ID *****

Abstract

Multimodal Large Language Models (MLLMs) excel at simple vision-language tasks but struggle when faced with complex tasks that require compositional capabilities, such as combining foundational capabilities like object recognition, spatial understanding, and counting. Visual Instruction Tuning (VIT), a critical training step for MLLMs, has traditionally focused on scaling data volume but overlooks the compositional complexity of training examples, limiting their effectiveness in real-world scenarios. We propose COMPACT, *COMPositional Atomic-to-complex Visual Capability Tuning that enables MLLMs to solve complex tasks by explicitly training them on compositions of foundational atomic capabilities. By generating training data with controlled compositional complexity and balanced distribution,* COMPACT *enables MLLMs to learn complex capabilities ($k \geq 1$) more efficiently. With only 10% of the* LLaVA-665K *training data,* COMPACT *achieves 100.18% of the performance obtained using the full dataset. We observe that training with* COMPACT *on questions requiring up to $k \leq 3$ capabilities exhibits strong generalization to complex multi-capability questions with $k > 3$ capabilities.* COMPACT *offers a scalable, data-efficient, atomic-to-complex visual compositional tuning recipe to improve on complex visual-language tasks.*

1. Introduction

Multimodal Large Language Models (MLLMs) like LLaVA [18, 19], Cambrian [30] and Eagle [17, 27] achieve impressive performance in general-domain visual-language tasks. However, in complex visual reasoning, solving a task often requires multiple foundational capabilities. Consider the following question: "Are there more blue squares or red circles on the image?" Many state-of-the-art models fail on such *compositional* questions, even though they can answer simpler questions correctly (e.g. "What color is the square?"). Such failures suggest that current models do not systematically generalize to questions with higher compositional complexity.

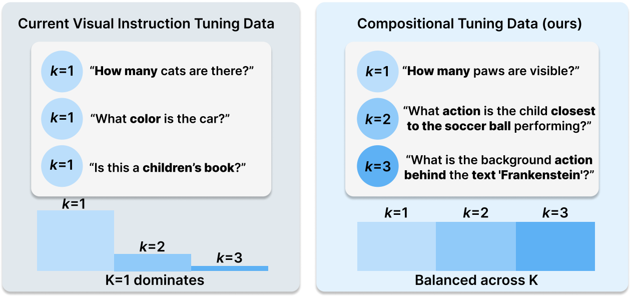
Current Visual Instruction Tuning Data
k=1 "How many cats are there?"
k=1 "What color is the car?"
k=1 "Is this a children's book?"
k=1
k=2
k=3
K=1 dominates
Compositional Tuning Data (ours)
k=1 "How many paws are visible?"
k=2 "What action is the child closest to the soccer ball performing?"
k=3 "What is the background action behind the text 'Frankenstein'?"
k=1
k=2
k=3
Balanced across K

Figure 1. **Compositional Complexity Comparison.** Comparison between visual instruction tuning data (LLaVA-665K) and our visual compositional tuning data (COMPACT). The previous VIT data is dominated by simple queries ($k = 1$), while our method ensures a balanced distribution across different levels of compositional complexity ($k = 1$ to $k = 3$).

Recent efforts to improve MLLMs' capabilities have mainly relied on scaling the Visual Instruction Tuning (VIT) data [17, 20, 21, 27, 30]. However, such datasets (e.g. LLaVA-665K [20]) are dominated by simple queries requiring only one capability, lacking sufficient *compositional complexity* (Fig. 1). While effective for simple tasks, recent studies have shown that even state-of-the-art MLLMs struggle with integrating capabilities and generalizing to complex visual tasks due to limitations in the compositional complexity of their training data [25, 33].

We introduce **COMPACT** (COMPositional Atomic-to-Complex Visual Capability Tuning), a method that scales capabilities of MLLMs from atomic ($k = 1$) to composite ($k > 1$) complexity levels. We combine atomic capabilities–object recognition, action recognition, spatial recognition, text recognition, color attribution, shape attribution, counting, spatial relationship understanding, and object interaction understanding–to generate a compositional training dataset that can promote a model's internalization of compositional structures of complex tasks. We summarize our key contributions:

- We introduce COMPACT, an effective visual compositional tuning approach that builds complex compositional capabilities from simple atomic capabilities. By systematically combining 10 foundational capabilities into training data of controlled complexity, COMPACT addresses a key limitation of conventional VIT methods that rely on incidental capability composition through data scaling.

- We develop a structured data recipe that covers a wider range of task regimes by enforcing a balanced distribution across different levels of compositional complexity ($k = 1, 2, 3$). This approach addresses the *complexity cliff* present in standard VIT datasets [20], where 90.1% of questions require two or fewer capabilities, limiting models' ability to handle more complex reasoning tasks.
- With just 32K samples of our compositional tuning data combined with 5% of LLaVA-665K (totaling only 10% of the original dataset size), COMPACT matches the full-scale baseline's performance (100.18% relative score) while demonstrating exceptional generalization to complex tasks. It shows remarkable gains on higher-complexity visual reasoning, improving performance by 49.8% on MMStar and 69.2% on MMVet for $k = 4$ tasks compared to the original data recipe.

2. Method

2.1. Atomic Visual Capabilities

Atomic (or basis) capabilities are foundational skills that can be combined to solve complex tasks. For example, a model needs to acquire object recognition, color attribution, and spatial relationship capabilities to identify how objects of different colors are spatially oriented. For each visual reasoning task T, we identify a set of atomic capabilities $\{c_1, \dots c_k\}$ required to solve this task. We define the number of atomic capabilities k required to solve the task T as its *compositional complexity*.

We build a taxonomy of atomic capabilities from existing literature on MLLMs and general visual tasks [13, 33]. Extremely low-frequency and non-perceptual capabilities (e.g. cultural knowledge, historical context, and math) are filtered, resulting in 10 fine-grained atomic capabilities (Tab. 3) that focus on visual understanding. We categorize these atomic capabilities into three major categories: **Attribution**, **Recognition**, and **Relation**.

2.2. Visual Compositional Tuning Data Recipe

In our proposed approach COMPACT, we generate multi-capability questions $\mathcal{D}_{\text{comp}}$ by prompting vision-language models to create questions that require natural [1] integration of exactly k atomic visual capabilities. This process involves four key steps:

Step 1: Capability Sampling. For each image in the training dataset, we randomly sample multiple rounds of $k \in \{1, 2, 3\}$ atomic capabilities from our predefined pool of 10 visual capabilities. At each iteration of capability sampling for an image, we keep track of the capabilities that have been chosen so far to prioritize the remaining ones. Duplicate combinations of capabilities for the same image are automatically dropped.

Step 2: Conversation Generation. For each capability combination that is sampled, we prompt Gemini-2.0-Flash to generate a conversational question-answer pair that integrates all capabilities in the combination, as well as a score between 0 and 100 that represents its confidence in output quality. Our carefully designed prompt (see Appendix C) enforces several key constraints: (1) questions must require examining the image and cannot be answered from its text alone, (2) answers must be concise, (3) questions must integrate exactly the specified capabilities naturally without conjunctions, and (4) questions must reference objects and features actually present in the image. The purpose of these constraints is to produce vision-centric conversations that are unambiguous and natural.

Step 3: Quality Verification. We include a verification process with Gemini-2.0-Flash to ensure the quality and diversity of the training dataset. We filter out questions that (1) have uninformative answers (e.g., "unknown", "not visible"), (2) have confidence scores below 70%, (3) share more than 60% of their words with previously accepted questions, or (4) can be answered from the question alone.

Then, we perform capability verification by prompting Gemini-2.0-Flash [29] to analyze whether each question requires exactly the k specified capabilities. Questions that additionally require unspecified capabilities or fail to utilize all specified capabilities are rejected. The generation and verification steps are repeated iteratively until we collect 2-3 high-quality questions per k for each image or reach a maximum of 10 verification attempts. Only images with at least two verified questions are included in the final dataset.

Step 4: Dataset Assembly. The final training dataset combines two components: (1) compositional tuning data–randomly sampled images from the LLaVA-665K dataset and their COMPACT-generated multi-turn conversations–and (2) instruction tuning data–random 5% subset of LLaVA-665K VIT dataset. This careful mixture serves a dual purpose. First, our compositional data improves the model's capability to reason about multiple visual aspects within a single question. Second, the VIT subset maintains the model's ability to handle diverse response formats required by modern MLLM benchmarks (e.g., multiple-choice questions [7], open-ended answers [20]). In this way, we delegate the instruction following capability training to the original VIT dataset, while allowing our compositional data to focus on developing compositional reasoning.

[1] We use the term "natural" to denote combination of visual capabilities that correspond to their co-occurrence patterns in real-world settings, wherein multiple capabilities are integrated in a way that is contextually and semantically meaningful.

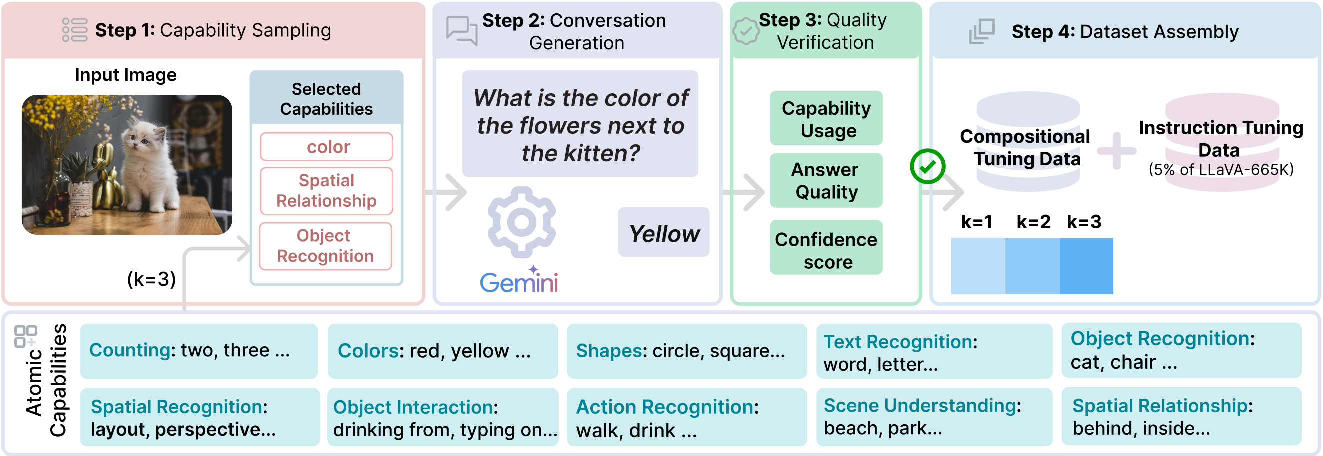
Step 1: Capability Sampling
Input Image
Selected Capabilities
color
Spatial Relationship
Object Recognition
(k=3)
Step 2: Conversation Generation
What is the color of the flowers next to the kitten?
Gemini
Yellow
Step 3: Quality Verification
Capability Usage
Answer Quality
Confidence score
Step 4: Dataset Assembly
Compositional Tuning Data
Instruction Tuning Data
(5% of LLaVA-665K)
k=1
k=2
k=3
Atomic Capabilities
Counting: two, three ...
Colors: red, yellow ...
Shapes: circle, square...
Text Recognition: word, letter...
Object Recognition: cat, chair ...
Spatial Recognition: layout, perspective...
Object Interaction: drinking from, typing on...
Action Recognition: walk, drink ...
Scene Understanding: beach, park...
Spatial Relationship: behind, inside...

Figure 2. **COMPACT**. *(Left):* We sample atomic capabilities ($k = 1$) such as color, object recognition, and spatial relationship. *(Center):* Based on the sampled capabilities, we generate questions ($k = 1, 2, 3$) with the required number of compositions. *(Right):* We verify the quality of the generated conversations and assemble them with instruction tuning data for instruction following capability. This structured data recipe explicitly models atomic-to-complex learning procedure, in contrast to standard VIT [20] dominated by simple queries.

Data	InfoVQA [23]	SeedBench [15]	MME [7]	TextVQA [28]	MMVet [36]	CV-Bench [30]	MMStar [4]	LLaVA-W [20]	Rel. (%)
LLAVA-665K	20.80	41.72	**1478.48**	**46.99**	29.22	**60.92**	35.11	**68.50**	100.00
Random	20.05	41.85	1327.70	42.88	30.46	54.71	34.13	64.30	95.38
ICONS [32]	21.0	42.03	1402.75	43.12	31.23	55.96	35.96	61.8	97.47
COMPACT (ours)	**23.68**	**43.13**	1379.94	44.37	**31.74**	55.28	**36.13**	64.50	100.18
$\bar{k}$	0.34	1.11	1.16	1.19	1.24	1.33	1.40	3.05	

Table 1. **Baseline Comparisons.** Performance comparison of COMPACT and multiple baselines. COMPACT integrates atomic capabilities into tasks of higher compositional complexity, enabling models to generalize and handle complex tasks without explicit decomposition. With 5% of LLAVA-665K instruction tuning data and our 32K compositional tuning data, COMPACT consistently outperforms standard VIT model trained on same amount of data as well as the LLAVA-665K trained model on multimodal benchmarks.

3. Experiments

3.1. Evaluation Testbed

Model. We train the LLaVA-v1.5-7B-LoRA [20] model–a post feature alignment checkpoint that has not been exposed to any visual instruction tuning data–on our COMPACT training dataset for one epoch with its official LoRA fine-tuning settings. The COMPACT training dataset includes 32K compositional samples and 5% of LLAVA-665K [20].

Baselines. We compare with the following baselines: **LLAVA-665K**: The standard LLaVA-1.5 model trained on the full LLAVA-665K dataset (665K samples), representing the conventional visual instruction tuning (VIT) paradigm. This serves as the primary performance baseline. **Random**: A random 10% subset of LLAVA-665K VIT dataset composed 65K samples, matching the size of our COMPACT training dataset. **ICONS** [32]: A gradient-driven influence consensus approach that selects a compact training dataset for data-efficient visual instruction tuning. This method uses influence functions to identify the most informative 65K samples across multiple tasks, which also matches the size of our COMPACT training dataset.

3.2. Main Results

Overall Performance. As shown in Tab. 1, COMPACT exceeds the LLAVA-665K baseline using only 5% of its data and 32K compositional tuning data. COMPACT outperforms both the random baseline [20] and ICONS [32] on most benchmarks, demonstrating superior generalization on multi-capability tasks.

We use Gemini-2.0-Flash to analyze each question and identify the atomic capabilities required to give an answer (see the details of the system prompt in §C). We average these numbers for each benchmark to compute benchmark-specific $\bar{k}$ values. COMPACT shows consistent improvements across different $\bar{k}$, achieving strong gains on tasks like InfoVQA (+13.8% over LLAVA-665K), Seed-Bench (+3.4%), MM-Vet (+8.6%), and MMStar (+2.9%) while maintaining competitive performance on TextVQA and LLaVA-in-the-Wild.

Data Efficiency of Compositional Tuning. We investigate the data efficiency of COMPACT by observing how model

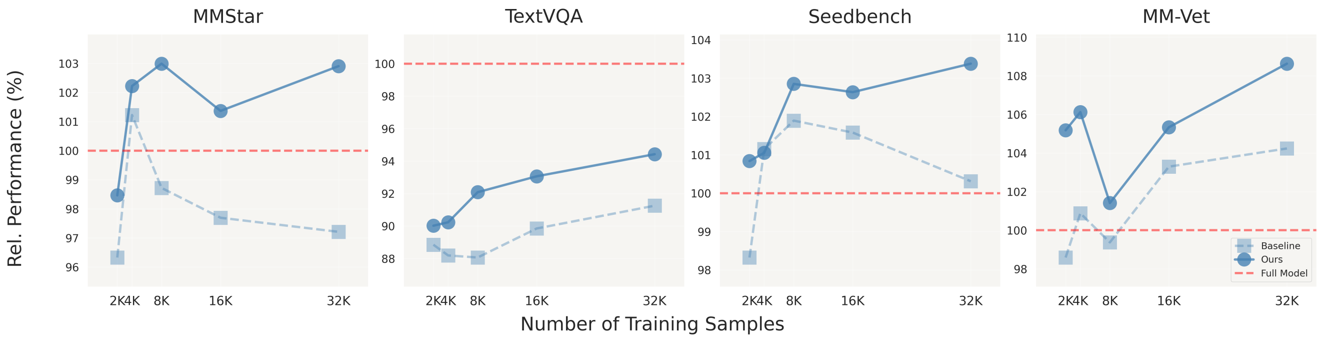

Figure 3. **Performance Comparison across Compositional Tuning Data Regimes.** We compare COMPACT (solid lines) with standard VIT (dashed lines) across benchmarks as the number of compositional tuning data increases from 2K to 32K. COMPACT consistently outperforms standard VIT across data regimes, achieving better performance with fewer data. The performance gap is pronounced for complex reasoning benchmarks such as MM-Vet and MMStar, where the 8K COMPACT model often matches or exceeds the standard VIT baseline at 32K. This demonstrates the data efficiency of COMPACT, requiring substantially less data than standard VIT to achieve comparable or better results.

performance changes with respect to the amount of compositional tuning data. Fig. 3 shows that COMPACT has a general upward trend across all benchmarks unlike the random baseline as the number of compositional tuning samples increases. Models trained on compositional tuning data often match or exceed the performance of LLaVA-665K and random baseline models trained on much larger data. We hypothesize that this improvement in data efficiency comes from two factors: (1) COMPACT continuously provides learning signals of higher compositional complexity by balancing the distribution of k. In contrast, the original VIT paradigm trains almost exclusively on $k = 1$ tasks (e.g., single-capability queries). Models trained on LLaVA-665K data receive signals of higher compositional complexity less frequently, leaving them unprepared for compositional generalization, the ability to integrate combinations of capabilities not explicitly seen during training. (more analysis in §D). (2) COMPACT sustains the learning potential during training by explicitly introducing diverse ($k \geq 1$) integrations of atomic capabilities. Meanwhile, original VIT relies heavily on simpler tasks ($k = 1$) that can be easily memorized and templatized for rapid saturation of learning potential.

Performance Gains on Complex Compositional Questions. COMPACT shows strong compositional generalization capabilities, achieving notable performance improvements on complex compositional questions. As shown in Tab. 2, our method achieves competitive performance across various levels of compositional complexity (k) on the MM-Vet and MMStar benchmarks. Despite not being explicitly trained on data with $k > 3$, our model effectively generalizes to higher complexity tasks. For instance, our method achieves a score of 55 for $k = 4$ and 20 for $k = 5$, while the model trained with LLaVA-665K, which is significantly larger, reaches 32.5 for $k = 4$ and 0 for $k = 5$. This shows that our method achieves robust performance in scenarios with higher compositional complexity.

Method	**k=0**	**k=1**	**k=2**	**k=3**	**k=4**	**k=5**
MM-Vet						
k-Distribution (%)	22.0	42.2	28.4	5.0	1.8	0.5
COMPACT (%)	7.0	**36.2**	47.4	**30.9**	**55.0**	**20.0**
LLaVA-665K VIT (%)	**7.8**	35.5	**52.5**	20.0	32.5	0.0
MMStar						
k-Distribution (%)	9.6	53.5	26.6	8.1	1.9	0.2
COMPACT (%)	27.1	**38.9**	**28.2**	**39.7**	**52.9**	0.0
LLaVA-665K VIT (%)	**32.9**	38.4	27.2	32.5	35.3	0.0

Table 2. **Compositional Generalization to Higher-Complexities.** Performance comparison across compositional complexities (k). COMPACT shows competitive performance against LLaVA-665K VIT training. It exceeds the LLaVA-665K baseline at higher compositional complexity tasks ($k = 4$ and $k = 5$) while using significantly less training data. The k-distribution rows show the distribution of compositional complexities in each benchmark.

4. Discussion

Conclusion. In this work, we introduce COMPACT, a structured data recipe that systematically combines atomic visual capabilities into composite capabilities to solve complex multimodal tasks. Our experimental results show that explicitly training with compositions of atomic capabilities achieves superior performance across benchmarks compared to standard VIT with only a small fraction of its data. Our work presents the potential of structured compositional learning as a scalable, data-efficient pathway toward multimodal models that can solve complex, multi-capability tasks via compositional generalization.

References

[1] Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above. *arXiv preprint arXiv:2502.14127*, 2025. 1

[2] Declan Campbell, Sunayana Rane, Tyler Giallanza, Camillo Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven Frankland, Tom Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems*, 37: 113436–113460, 2025. 1

[3] Hyunsik Chae, Seungwoo Yoon, Chaeyoung Chun, Gyehun Go, Yongin Cho, Gyeongmin Lee, and Ernest K Ryu. Decomposing complex visual comprehension into atomic visual skills for vision language models. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*. 1

[4] 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. 3, 1

[5] Mingyang Chen, Haoze Sun, Tianpeng Li, Fan Yang, Hao Liang, Keer Lu, Bin Cui, Wentao Zhang, Zenan Zhou, and Weipeng Chen. Facilitating multi-turn function calling for llms via compositional instruction tuning. *arXiv preprint arXiv:2410.12952*, 2024. 1

[6] Jerry A. Fodor and Ernest LePore, editors. *The Compositionality Papers*. Oxford University Press, 2002. 1

[7] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: a comprehensive evaluation benchmark for multimodal large language models. corr abs/2306.13394 (2023), 2023. 2, 3, 1

[8] Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, Dinesh Manocha, et al. A closer look at the limitations of instruction tuning. *arXiv preprint arXiv:2402.05119*, 2024. 1

[9] Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*, 2024. 1

[10] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023. 1

[11] Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*, 2024. 1

[12] Jiaxing Huang, Jingyi Zhang, Kai Jiang, Han Qiu, and Shijian Lu. Visual instruction tuning towards general-purpose multimodal model: A survey. *arXiv preprint arXiv:2312.16602*, 2023. 1

[13] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 2

[14] Jaewoo Lee, Boyang Li, and Sung Ju Hwang. Concept-skill transferability-based data selection for large vision-language models. *arXiv preprint arXiv:2406.10995*, 2024. 1

[15] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 3, 1

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1

[17] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025. 1

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 2, 3

[21] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1

[22] Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. Less is more: High-value data selection for visual instruction tuning. *arXiv preprint arXiv:2403.09559*, 2024. 1

[23] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 3, 2

[24] Timothy Ossowski, Ming Jiang, and Junjie Hu. Prompting large vision-language models for compositional reasoning. *arXiv preprint arXiv:2401.11337*, 2024. 1

[25] Simon Park, Abhishek Panigrahi, Yun Cheng, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Generalizing from simple to hard visual reasoning: Can we mitigate modality imbalance in vlms? *arXiv preprint arXiv:2501.02669*, 2025. 1

[26] Eva Sánchez Salido, Julio Gonzalo, and Guillermo Marco. None of the others: a general technique to distinguish reasoning from memorization in multiple-choice llm evaluation benchmarks. *arXiv preprint arXiv:2502.12896*, 2025. 1

[27] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024. 1

[28] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 3, 2

[29] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[30] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2025. 1, 3, 2

[31] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 1

[32] Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. Icons: Influence consensus for vision-language data selection. *arXiv preprint arXiv:2501.00654*, 2024. 3, 1

[33] Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. *arXiv preprint arXiv:2408.14339*, 2024. 1, 2

[34] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*, 2023. 1

[35] Zhuoyan Xu, Zhenmei Shi, and Yingyu Liang. Do large language models have compositional ability? an investigation into limitations and scalability. *arXiv preprint arXiv:2407.15720*, 2024. 1

[36] 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. *arXiv preprint arXiv:2308.02490*, 2023. 3, 1

[37] Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. *Advances in neural information processing systems*, 35:29776–29788, 2022. 1

[38] Haoyu Zhao, Simran Kaur, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Can models learn skill composition from examples? *Advances in Neural Information Processing Systems*, 37:102393–102427, 2024. 1

[39] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. 1