COMPACT: COMPositional Atomic-to-Complex Visual Capability Tuning

Xindi Wu[1*] Hee Seung Hwang[1*] Polina Kirichenko[2] Olga Russakovsky[1]

[1]Princeton University [2]Meta AI

https://princetonvisualai.github.io/compact/

Abstract

Multimodal Large Language Models (MLLMs) excel at simple vision-language tasks but struggle when faced with complex tasks that require multiple capabilities, such as simultaneously recognizing objects, counting them, and understanding their spatial relationships. This might be partially the result of the fact that Visual Instruction Tuning (VIT), a critical training step for MLLMs, has traditionally focused on scaling data volume, but not the compositional complexity of training examples. We propose COMPACT *(COMPositional Atomic-to-Complex Visual Capability Tuning), which generates a training dataset explicitly controlling for the compositional complexity of the training examples. The data from* COMPACT *allows MLLMs to train on combinations of atomic capabilities to learn complex capabilities more efficiently. Across all benchmarks,* COMPACT *achieves comparable performance to the* LLaVA-665K *VIT while using less than 10% of its data budget, and even outperforms it on several, especially those involving complex multi-capability tasks. For example,* COMPACT *achieves substantial 83.3% improvement on MMStar and 94.0% improvement on MM-Vet compared to the full-scale VIT on particularly complex questions that require four or more atomic capabilities.* COMPACT *offers a scalable, data-efficient, visual compositional tuning recipe to improve on complex visual-language tasks.*

1. Introduction

Multimodal Large Language Models (MLLMs) have shown impressive progress in a wide range of vision-language tasks [1, 2, 7]. Yet, from early diagnostic works [15, 16] to recent state-of-the-art models like LLaVA [9, 10], Cambrian [20], and Eagle [8, 17], compositionality remains a challenge. Consider the following question: "Are there more blue squares or red circles on the image?" A model that is capable of recognizing shapes, colors and counting objects should be able to answer it correctly. Despite years of progress, state-of-the-art models still fail on such *compositional* questions, even though they can answer simpler ones correctly (e.g. "What color is the square?"). This has been a long-standing issue and such failures suggest that current models do not systematically generalize to tasks with higher compositional complexity.

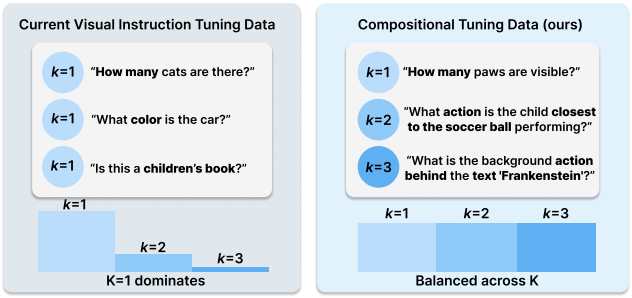

Figure 1. **Compositional Complexity Comparison.** Comparison between visual instruction tuning data (LLaVA-665K [11] VIT) and our compositional tuning data (COMPACT). The VIT data is dominated by simple queries ($k = 1$), while our COMPACT data is balanced across compositional complexity levels ($k = 1, 2, 3$).

To address this, recent efforts have primarily scaled up the amount of training data used for Visual Instruction Tuning (VIT) [8, 11, 12, 17, 20], an essential but *data- and compute-heavy* step for MLLM training. However, such datasets (e.g. LLaVA-665K [11]) are dominated by simple queries that require only one capability, lacking sufficient *compositional complexity* (Fig. 1). Even with large-scale instruction tuning, recent studies show that models still struggle with integrating capabilities and generalizing to complex visual tasks due to limitations in the compositional complexity of their training data [14, 22].

Instead of treating compositionality as a byproduct of scale, we encourage compositional capabilities in MLLMs with hierarchically structured compositional training data. In this work, we introduce **COMPACT** (COMPositional Atomic-to-complex Visual Capability Tuning), a data

*Equal contribution

recipe that scales capabilities of MLLMs from atomic ($k = 1$) to composite ($k > 1$) complexity levels. We define a set of 10 atomic capabilities and then combine them to generate a compositional training dataset that can promote model's internalization of the compositional structures of complex tasks in a compute-efficient manner. We summarize our key contributions:

- We introduce COMPACT, a visual compositional tuning data recipe that builds complex capabilities from simple atomic capabilities. By systematically combining 10 atomic capabilities to control the complexity of training samples, 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 enforces a balanced distribution across different levels of compositional complexity ($k = 1, 2, 3$) to cover a wider range of task regimes. This approach flattens the *complexity cliff* in conventional VIT datasets [11], where 90.1% of the questions require two or fewer capabilities.

2. Method

2.1. Atomic Visual Capabilities

Atomic 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 understanding capabilities to identify how objects of different colors are spatially oriented. For each task T, we identify a set of atomic visual capabilities $\{c_1, \dots c_k\}$ required to solve this task. We define the number of atomic capabilities required to solve the task T as its *compositional complexity* k.

We build a taxonomy of atomic capabilities from the existing literature on MLLMs and their general visual reasoning skills [5, 22]. Extremely low-frequency and non-perceptual capabilities (e.g. cultural knowledge, historical context, and math) are removed, resulting in 10 fine-grained atomic capabilities that focus on visual understanding (Fig. 2). We categorize these atomic capabilities into three major categories: **Attribution** (color and shape), **Recognition** (objects, actions, text, spatial recognition, and counts), and **Relation** (spatial relationship, object interaction, or scene understanding).

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. We start by taking a random sample of images from LLAVA-665K [11]. For each image, we repeatedly sample $k \in \{1, 2, 3\}$ capabilities from our predefined pool of 10 atomic visual capabilities. We do the following in each round of capability sampling: (a) prioritize the capabilities that have not been selected for that image, and (b) drop duplicate combinations of capabilities for the same image. These efforts ensure that our training examples efficiently capture diverse visual information from the images.

Step 2: Conversation Generation. For each capability combination that is sampled, we prompt Gemini-2.0-Flash [19] 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 the quality of the conversation. Our carefully designed prompt enforces several key constraints: (a) questions must require the use of visual information from the image and cannot be answered from its text alone, (b) answers must be concise, (c) questions must integrate exactly the specified capabilities naturally (without using conjunctions to simply conjoin single-capability questions), and (d) 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 [19] to ensure the quality and diversity of the training dataset. We filter out questions with uninformative answers (e.g., "unknown", "not visible") or those with confidence scores below 70%. Then, we perform capability verification by prompting Gemini-2.0-Flash [19]. Questions that require unspecified capabilities or do not utilize all k capabilities are rejected. The generation and verification processes in steps 2 and 3 repeat iteratively until we collect 2-3 high-quality conversations per k for each image.

Step 4: Dataset Assembly. The final training dataset combines two components: (1) a random 5% subset of the LLAVA-665K [11] VIT dataset, and (2) our COMPACT-generated compositional tuning data. The VIT subset maintains the model's ability to handle diverse response formats and instructions required by modern MLLM benchmarks (e.g., multiple-choice questions [4], open-ended answers [11]). On the other hand, our compositional data trains the model's capability to reason about multiple visual aspects within a single complex question.

[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.

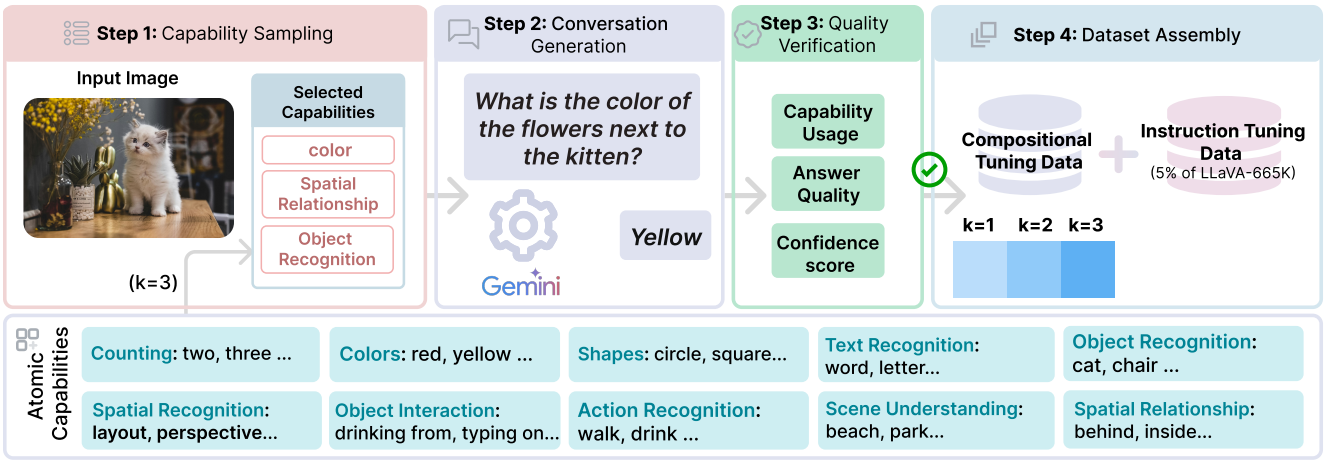

Figure 2. **COMPACT's Data Generation Pipeline.** *(Left):* We sample $k \in \{1, 2, 3\}$ atomic capabilities such as color, object recognition, and spatial relationship. *(Center):* We generate questions that integrate all k sampled capabilities. *(Right):* We verify the quality of generated conversations and combine them with instruction tuning data to maintain instruction following capability. This structured data recipe explicitly models atomic-to-complex learning procedure, in contrast to standard LLaVA-665K [11] VIT that promotes learning from simple queries.

Recipe	# Data	InfoVQA [13]	SeedBench2Plus [6]	MME [4]	TextVQA [18]	MM-Vet [23]	CV-Bench [20]	MMStar [3]	LLaVA-W [11]	Rel. (%)
LLaVA-665K [11]	665K	20.80	41.72	**1478.48**	**46.99**	29.22	**60.92**	35.11	**68.50**	100.00
Random	65K	20.05	41.85	1327.70	42.88	30.46	54.71	34.13	64.30	95.38
ICONS [21]	65K	21.0	42.03	1402.75	43.12	31.23	55.96	35.96	61.8	97.47
COMPACT (ours)	65K	**23.68**	**43.13**	1379.94	44.37	**31.74**	55.28	**36.13**	64.50	**100.18**

Table 1. **Baseline Comparisons.** Performance comparison of COMPACT with baselines. With only 5% of the LLaVA-665K [11] VIT data and 32K of our compositional tuning data (65K total), COMPACT outperforms the random subset of the VIT data (Random), gradient-based approach selected subset of the VIT data (ICONS [21]), and even the full VIT data on diverse multimodal benchmarks. The best and second best results for each benchmark are shown in **bold** and underlined, respectively. COMPACT integrates atomic capabilities into tasks of higher compositional complexity, enabling models to generalize and handle complex tasks without explicit decomposition.

3. Experiments

3.1. Evaluation Testbed

Model. We train LLaVA-v1.5-7B-LoRA [11] model's pre-visual-instruction-tuning checkpoint[2] on our COMPACT training dataset. This checkpoint has not been exposed to any visual instruction tuning data prior to COMPACT training. The training dataset includes 32K-sample compositional tuning data unless otherwise stated. Additionally, we mix 5% of LLaVA-665K [11] to preserve instruction following capability. We train the model for one epoch with its official LLaVA-v1.5 LoRA fine-tuning settings.

Baselines. We compare the effectiveness of our COMPACT data recipe with several baseline datasets by training models with the same architecture under identical training configurations. **LLaVA-665K**: The full LLaVA-665K [11] VIT dataset (665K samples) used in LLaVA-v1.5. This serves as our primary performance baseline. **Random**: A 65K-sample random subset of LLaVA-665K [11] that matches our training data size. This baseline controls for data volume. **ICONS** [21]: A 65K-sample subset of LLaVA-665K [11] selected using the ICONS method, which is a gradient-driven influence-consensus based data selection method that selects the most informative samples for data-efficient visual instruction tuning. We evaluate models trained with different data recipes on multimodal benchmarks that assess complex visual capabilities.

[2] LLaVA-v1.5-mlp2x-336px-pretrain-vicuna-7b-v1.5, which has no prior exposure to visual instruction tuning data.

3.2. Main Results

Overall Performance. As shown in Tab. 1, our COMPACT achieves an average relative performance of 100.18%, outperforming even the full LLaVA-665K [11]. In comparison, the random baseline achieves 95.38%, and the ICONS [21] baseline 97.47%, highlighting the effectiveness of our compositional data generation strategy.

Visual Compositional Tuning is Data-Efficient. We study the data efficiency of COMPACT by analyzing how

Figure 3. **Performance Across Compositional Tuning Data Scales.** We fix the VIT subset (5% of LLaVA-665K [11]) and scale the compositional tuning data in COMPACT from 2K to 32K. For comparison, we remove the compositional tuning data and add more VIT data (2K-32K) instead to prepare VIT only recipes with equal data budgets. COMPACT (solid lines) consistently outperforms LLaVA-665K [11] VIT (dashed lines) with fewer data. The performance gap is pronounced for complex reasoning benchmarks such as MM-Vet and MMStar, where the 8K COMPACT model often exceeds the LLaVA-665K [11] VIT baseline at 32K. This demonstrates the data efficiency of COMPACT, requiring substantially less data than LLaVA-665K [11] VIT to achieve comparable or better results.

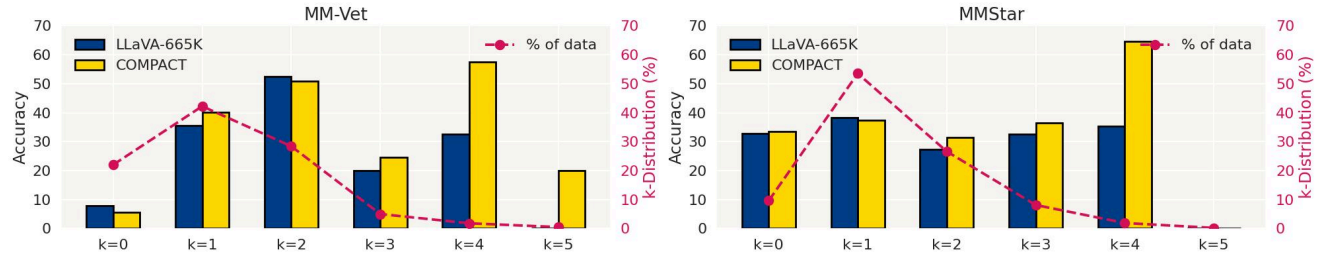

Figure 4. **Compositional Generalization to Higher-Complexities.** Performance comparison across compositional complexities (k). COMPACT shows competitive performance against LLaVA-665K [11] VIT training. It exceeds the LLaVA-665K [11] baseline at higher compositional complexity tasks ($k = 4$ and $k = 5$) while using significantly less training data. The k-distribution rows show the distribution of compositional complexities in each benchmark.

its performance changes as we scale the amount of compositional tuning data. We fix the VIT subset (5% of LLaVA-665K [11]) and scale the compositional tuning data in COMPACT from 2K to 32K. As comparison, we remove the compositional tuning data and add more VIT data (2K-32K) instead to match the dataset size. Fig. 3 shows that as the number of compositional tuning samples increases, COMPACT performance trends upward across all benchmarks while the random baseline shows mixed behavior as the size of the dataset increases. Furthermore, across all dataset sizes, COMPACT performs consistently better than the random baseline, and the gap increases as the size of the dataset grows. This demonstrates that COMPACT makes more effective use of training data compared to the baselines.

Performance Gains on Complex Compositional Questions. COMPACT's notable performance improvements on complex compositional questions demonstrate its potential for strong compositional generalization. As shown in Fig. 4, COMPACT achieves competitive performance on the MM-Vet [23] and MMStar [3] benchmarks across various levels of compositional complexity (k). Despite not being explicitly trained on $k > 3$ data, our model effectively generalizes to even higher k regimes. For MM-Vet [23], the scores are 57.5 (COMPACT) vs 32.5 (LLaVA-665K [11]) when $k = 4$, and 20.0 (COMPACT) vs 0.0 (LLaVA-665K [11]) when $k = 5$. For MMStar [3], the scores are 64.7 (COMPACT) vs 35.3 (LLaVA-665K [11]) when $k = 4$. This shows that COMPACT performs robustly in scenarios with higher compositional complexity.

4. Discussion

Conclusion. In this work, we introduce COMPACT, a data recipe that systematically combines atomic visual capabilities (e.g., object recognition, spatial reasoning, shape attribution) into composite capabilities to solve complex multimodal tasks. Our experimental results show that explicit training on compositions of atomic capabilities matches the full LLaVA-665K [11] VIT in performance across benchmarks with less than 10% of its data budget. 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] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2, 2023. 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. 3, 4

[4] 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

[5] 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

[6] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 3

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[8] 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

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

[10] 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

[11] 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, 4

[12] 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

[13] 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

[14] 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

[15] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. *arXiv preprint arXiv:1909.04696*, 2019. 1

[16] Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10003–10011, 2020. 1

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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, 4