COMPACT: COMPositional Atomic-to-Complex Visual Capability Tuning

Supplementary Material

In this supplementary material, we provide related work (§A), additional method details (§B), additional experiment details (§C), additional analysis (§D), ablations (§E), and additional discussions (§F). Finally we include visualizations (§G) that demonstrate the effectiveness of compositional tuning through comparative case studies.

A. Related Work

Visual Instruction Tuning. Instruction following is an essential capability in language models [38, 46]. Misalignment between a model's response and the format requested by a question can hinder the precise evaluation of its performance and capabilities [3, 11, 12, 31]. In order to adapt MLLMs to respond appropriately to diverse question formats (e.g., multiple-choice, short- and long-response questions), Visual instruction tuning [22, 24] has been proposed. VIT involves training a model on a fixed set of instruction patterns that can be repeated during inference. Although VIT has shown performance improvements in general multimodal capabilities [14], recent work [10] has shown that optimizing for response formatting potentially limits the quality of language model responses.

While VIT [24] focuses on learning simple capabilities via instruction following, our data recipe explicitly models compositional capabilities in the training data. Our approach directly addresses the lack of exposure to compositional questions during training, enabling models to improve on tasks that are more complex in capability space. COMPACT leverages the learning potential of both compositional tuning and instruction tuning data to create a more optimal data recipe than conventional VIT.

Compositionality in LLMs and MLLMs. Semantically, compositionality is the claim that the meaning of a complex statement is a result of the combination of its constituents. [8]. In the context of visual capabilities of MLLMs, compositional capability refers to a model's ability to perform complex tasks by combining multiple capabilities [13], where each capability is related to understanding basic visual concepts such as objects, attributes, relationships. Recent work has shown that compositional capability can be trained in LLMs [42, 45], but generalizations to the realms of MLLMs have been largely incomplete. Some studies highlight that while MLLMs do show signs of compositional capability [28], they struggle when constituting components and their combined patterns are not strongly learned or missing during training [4]. Furthermore, previous works have focused on limited domains such as geometry [5], visual recognition, and language [7], or employed a relaxed definition of compositionality as a sequential array of tasks [20] rather than integrating them.

Studies show that general visual capability requires strong compositional ability [44]. In order to train MLLMs to learn complex capabilities, it is necessary to explicitly model compositionality in the training data. Our approach takes advantage of these findings to create a data recipe for training complex capabilities across visual domains.

Data Efficiency in MLLMs. VIT is a data and compute-heavy step in training [41]. Studies have found that the performance of MLLMs can be reproduced with less data and better techniques, suggesting that the amount of data needed for VIT can be reduced. For example, recent works have developed effective VIT data recipes by leveraging data selection methods and curating higher-quality training datasets [16, 26]. ICONS [39] shows that models can achieve near-perfect performance across a suite of MLLM benchmarks with a fraction of the original VIT dataset. On the other hand, some studies proposed an alternative approach of scaling up to improve visual capabilities even further [18].

However, these approaches treat compositionality as a byproduct of scale rather than as a learnable capability. Our COMPACT formalizes atomic capabilities and systematically incorporates their combinations into the training dataset to efficiently address the limitations in generalization to complex compositional tasks. By redistributing the compositional complexity of the training data, we scale the model's exposure to complex tasks without scaling the data.

B. Additional Method Details

Tab. 2 shows the taxonomy of atomic capabilities and their detailed descriptions.

COMPACT preserves the contents of the images sampled from LLaVA-665K [24] when generating new multi-turn conversations. This enables us to fairly compare COMPACT and existing methods in their ability to extract rich and structured information from the controlled set of images. We further adjust the ratio of the VIT subset to our compositional tuning data and study how it affects performance (§E). Our findings show the optimal balance between the preservation of instruction following capability and the training of compositional capabilities.

C. Additional Experiment Details

Benchmarks. We evaluate models trained with different data recipes on established multimodal benchmarks that assess complex visual capabilities. 1) **MM-Vet** [43] includes

Table 2. **Taxonomy of Atomic Capabilities.** We identify 10 atomic capabilities and categorize them into three groups: Attribution, Recognition, and Relation. Atomic capabilities serve as building blocks for compositional instruction tuning. For each capability, we provide the definition and a question example that requires the capability to answer.

Group	Capability	Definition	Example Question
Attribution	Color	Identifying or comparing colors of objects in the image	What **color** is the car?
	Shape	Recognizing and describing the shapes of objects in the image	What **shape** is the dining table?
Recognition	Object Recognition	Identifying and naming objects present in the image	What **object** is on the table?
	Action Recognition	Identifying what action is being performed	What is the person **doing** in this image?
	Text Recognition	Reading and interpreting text visible in the image	What **word** is written on the sign?
	Spatial Recognition	Understanding the overall spatial layout and arrangement of the entire scene	How is the furniture **arranged** in this room?
	Counting	Determining the number of instances of something in the image	**How many** people are in the room?
Relation	Spatial Relationship	Identifying how specific objects are positioned relative to each other	What is **next to** the red car?
	Object Interaction	Analyzing how multiple objects interact with each other	How is the woman **interacting** with the laptop?
	Scene Understanding	Identifying the type of environment/setting	**Where** is this scene taking place?

16 types of complex multimodal tasks integrated from 6 core capabilities (recognition, OCR, knowledge, language generation, spatial awareness, and math). 2) **MME** [9] contains 10 perception (e.g., color, count, OCR) and 4 cognition (e.g., commonsense reasoning, text translation, code understanding) related visual subtasks. 3) **LLaVA-in-the-Wild** [24] is an open-ended visual question answering benchmark that asks complex questions on real-world images. 4) **SeedBench2Plus** [17] evaluates visual comprehension skills of MLLMs with a focus on charts, maps, and webs. 5) **MMStar** [6] contains 1,500 visual questions that span 6 core capabilities (fine-grained perception, coarse perception, mathematics, science & technology, logical reasoning and instance reasoning), carefully curated to evaluate multimodal understanding. 6) **CV-Bench** [37] is a MLLM benchmark specialized for 2D and 3D visual understanding that includes spatial relationship, object count, relative distance, and depth order. 7) **TextVQA** [35] evaluates visual understanding of texts in the image. 8) **InfoVQA** [27] measures visual understanding of infographic images. These benchmarks cover a broad range of vision-centric capabilities. We also note that some of these benchmarks include non-visual questions involving skills as knowledge and math, which are not our primary focus. We provide a more detailed discussion of model performance in these knowledge-intensive and math-intensive tasks in Appendix §D.

System prompts. We provide the system prompt for our capability analysis where we identify all the required capabilities for a given question **(A)**. We also provide the system prompts for compositional question generation **(B)** and verification **(C)**. The generation prompt includes structured guidelines to ensure that the generated multi-capability questions naturally blend different capabilities and can only be answered by checking the corresponding images. The verification prompt checks if the questions meet these guidelines and do not contain subjective interpretations or compositional flaws.

Example Questions with Different Compositional Complexities

MM-Vet ($k = 3$):
Q: What is the color of the hat worn by the person in the front left?
Required capabilities: color attribution, object recognition, spatial relationship
MMStar ($k = 4$):
Q: What is the position of the red rug in the living room?
Required capabilities: color attribution, object recognition, spatial relationship, scene understanding
MMStar ($k = 5$):
Q: Is the number of metal cars that are left of the tiny matte school bus greater than the number of tiny cyan double bus?
Required capabilities: spatial relationship, object recognition, counting, color attribution, shape attribution

(A) System Prompt for Capability Analysis

Prompt: You are an AI assistant that analyzes questions to identify the core capabilities required to answer them. Given a question, identify ALL the capabilities it requires from this list:
- spatial relationship (understanding relative positions)
- object interaction (how objects/people interact)
- object relationship (relationships between objects)
- text recognition (reading text in images)
- spatial recognition (understanding 3D space)
- action recognition (identifying actions/activities)
- object recognition (identifying objects)
- counting (counting objects/people)
- color (identifying colors)
- shape (identifying shapes)

Return ONLY a JSON array of the required capabilities, like: [“capability1”, “capability2”]

(C) System Prompt for Question Verification

Prompt: You are an AI assistant that verifies if questions about images properly utilize specified capabilities.
Given a question and its answer, analyze whether it NATURALLY requires using EXACTLY k specified capabilities - no more, no less.
IMPORTANT:

- The question should require ALL specified capabilities to be answered
- The question should not require additional major capabilities beyond those specified
- The capabilities must be naturally integrated, not artificially forced

D. Additional Analysis

Distribution of Visual Capabilities. We use Gemini-2.0-Flash [36] to analyze each question and identify the atomic capabilities required to give an answer (see the details of the system prompt in Appendix §C). Fig. 5 shows the approximate distribution of the number of capabilities required per question in the LLaVA-665K [24] VIT dataset. We sampled 5,668 questions that belong to 1,000 random data points in the VIT dataset, and analyzed their compositional complexity using Gemini-2.0-Flash [36]. The mean compositional complexity of the questions is approximately $k = 1.5$, and the mode is $k = 1$. 59.2% of the questions utilize only one capability, and an additional 30.9% use 2 capabilities. Together, about 90% of the questions require 2 or less visual atomic capabilities. This *complexity cliff* in the LLaVA-665K [24] VIT dataset characterized by the scarcity of higher k questions leads to steep declines in its downstream performance on higher k tasks. (Fig. 4). The performance gap between COMPACT and the VIT for different k values shows that the VIT's complexity-agnostic training leaves models unprepared for tasks that require compositional generalization.

Interestingly, a small fraction of the questions (0.2%) require as many as 10 capabilities (e.g., "Question: Describe this photo in detail."). We also observe that 1.1% of the questions require zero capabilities, as illustrated in Fig. 5. We further provide $k = 0$ examples in Appendix §G. Fig. 6 shows the relative frequencies of atomic capabilities in the question samples. Object recognition (38.97%) and scene understanding (28.58%) are the most common. Other notable capabilities include spatial relationship (25.14%), text recognition (24.68%), and color attribution (14.40%). Less frequent capabilities include object interaction (6.55%), action recognition (6.05%), counting (2.95%), shape attribution (1.13%), and spatial recognition (1.06%).

Analysis of Conversation Length Distribution in LLaVA-665K. Fig. 7 shows the distribution of the number of conversations per image in LLaVA-665K [24]. 93.6% of the samples fall below the 10-pair threshold. The distribution's mean of 5.18 conversations per image ($\sigma = 5.62$) shows that the data is heavily skewed towards lower values. We fix the target number of conversations per image in the compositional tuning dataset based on these findings. We ensure a fair comparison by aligning the distribution of our data with the baseline distribution.

Table 3. **Limited Performance Improvements on Knowledge-Intensive Benchmarks.** Comparison shows modest improvements over random baseline on tasks that require substantial world knowledge or domain expertise. Numbers reported in accuracy (%) and relative performance to full model (%).

Model	OK-VQA	MMMU	MMMU-Pro		**Rel.**
			Standard	Vision	(Avg.)
Random	49.30	32.89	18.15	11.44	92.0%
COMPACT	50.02	33.89	20.23	11.91	96.6%
LLaVA-665K [24]	57.96	33.89	20.12	11.97	100%

Analysis of Limited Performance Gains on Knowledge-Intensive Tasks. While our compositional tuning approach shows general improvements on various benchmarks, we observe more modest gains in knowledge-intensive tasks. Table 3 compares the performance of different approaches on OK-VQA, MMMU, and MMMU-Pro benchmarks. COMPACT with 32k compositional tuning data shows relatively small improvements over the random baseline: OK-VQA (50.02% vs 49.30%), MMMU (33.89% vs 32.89%), and MMMU-Pro (20.23% vs 18.15% on standard tasks, 11.91% vs 11.44% on vision tasks). Notably, training on the full LLaVA-665K [24] VIT dataset leads to limited performance improvements on MMMU (33.89%). Although knowledge-related tasks are not our main focus, this inspires future work on designing compositional tuning approaches that cover broader capabilities outside of the vision space.

E. Ablations

We conduct a series of ablation studies to investigate key design considerations (compositional complexity distribution, atomic capability coverage, compositional complexity range, and instruction tuning ratio) in COMPACT. Unless otherwise specified, all experiments use 5% of LLaVA-665K [24] VIT data and 16K $k = 1, 2, 3$ compositional tuning data.

Effect of Matching LLaVA-665K Distribution. In order to show that the performance improvement of COMPACT mainly comes from the balanced distribution of compositional complexity in the compositional tuning data, we analyze the impact on performance when its compositional complexity is unbalanced. We further generate a 16K-

(B) System Prompt for Question Generation

Prompt: You are an AI assistant that generates challenging but well-defined questions and answers about images. First, I will provide you with k specific capabilities. Generate 1 question that naturally integrates EXACTLY these k capabilities.

IMPORTANT:

- If the question can be answered without looking at the image (e.g., the answer can be inferred from the question itself or previous questions), it's a BAD question
- Questions should be reasonably challenging but must have clear, unambiguous answers
- All answers must be extremely concise - use only a single word or short phrase
- Each question must be a single, integrated question that naturally combines all k given capabilities
- DO NOT use "and" or commas to combine separate questions
- Questions should require careful observation and reasoning
- Only generate questions when you can determine the answer with high confidence
- Avoid subjective or ambiguous questions
- ONLY ask about objects and capabilities that are ACTUALLY PRESENT in the image
- NEVER create questions about objects or features that don't exist in the image
- Generate diverse questions that differ in topic and required reasoning

CAPABILITY DEFINITIONS:

- spatial_relationship: Identifying how specific objects are positioned relative to each other (above, below, next to, inside, etc.) - focuses on the direct relationship between two or more particular objects
- spatial_recognition: Understanding the overall spatial layout and arrangement of the entire scene - focuses on the general organization, depth, perspective, or environmental context, rather than relationships between specific objects
- text_recognition: Reading and interpreting text visible in the image
- action_recognition: Identifying what action is being performed (can involve a single person/object)
- object_interaction: Analyzing how multiple objects interact with each other (requires at least two objects) - MUST involve at least one moving/active object, not just static objects positioned together - can include humans interacting with objects and humans interacting with humans
- object_recognition: Identifying and naming objects present in the image
- counting: Determining the number of instances of something in the image
- color: Identifying or comparing colors of objects in the image
- shape: Recognizing and describing the shapes of objects in the image
- scene_understanding: Identifying where the image is taken or the type of environment/setting (indoor/outdoor, beach, mountain, kitchen, office, etc.) - focuses on identifying the overall scene, background, or context of the image

Examples:

- **BAD**: "What color is the car, and where is it located?" (two separate questions)
- **BAD**: "What might the person be thinking?" (subjective/ambiguous)
- **BAD**: "Is this a nice room?" (subjective)
- **BAD**: "What breed of dog is in the corner?" (when no dog exists in the image)
- **BAD**: "How are the fridge and desk interacting?" (static objects don't qualify as interaction)
- **BAD**: "What is the color of the red car?" (answer can be inferred from the question itself without seeing the image)
- **GOOD**: "What color car is parked next to the red brick building?" (specific, clear answer)
- **GOOD**: "How many yellow tennis balls are visible on the wooden court?" (requires counting + color)
- **GOOD**: "What is the person in blue using to interact with the television?" (proper object interaction)
- **GOOD**: "Where is this image taken?" (scene understanding)
- **GOOD**: "Where is this scene happening?" (scene understanding)

sample compositional tuning data whose distribution of k resembles that of LLaVA-665K [24], which is heavily skewed as in Fig. 5. This gives us 58,168 $k = 1$, 30,364 $k = 2$, and 7,468 $k = 3$ conversations. Similar to the original COMPACT data recipe, we mix the unbalanced 16K compositional tuning data with the random 5% subset of the VIT data. We compare this unbalanced COMPACT training dataset with the following baselines: 1) a same size random subset of the VIT data, 2) the original COMPACT with 16K balanced compositional tuning data, and 3) the full VIT dataset. As shown in Tab. 4, the performance of unbalanced COMPACT stands at 96.62% (in relation to the full baseline), close to the random baseline at 96.28%. However, the performance of original COMPACT jumps to 98.83%, suggesting that most of the performance gain in COMPACT comes from the fair representation of higher k samples in the compositional tuning data.

Recipe	#Data	InfoVQA [27]	SeedBench2Plus [17]	MME [9]	TextVQA [35]	MMVet [43]	CV-Bench [37]	MMStar [6]	LLaVA-W [24]	Rel. (%)
LLaVA-665K [24]	665K	20.80	41.72	1478.48	46.99	29.22	60.92	35.11	68.50	100.00
Random	49K	20.33	42.38	1290.45	42.22	30.18	54.75	34.3	70.5	96.28
Unbalanced COMPACT	49K	22.28	41.17	1339.24	43.08	29.22	55.84	34.8	64.5	96.62
COMPACT	49K	22.68	42.82	1362.68	43.73	30.78	54.69	35.59	66.6	98.83

Table 4. **Matching LLaVA-665K Distribution.** Performance comparison of unbalanced COMPACT and multiple baselines. The distribution of compositional complexity in unbalanced COMPACT follows LLaVA-665K [24]. Training a model on unbalanced COMPACT leads to performance on par with training on the random baseline which is a subset of LLaVA-665K [24] equal in size, suggesting that a balanced distribution of k in compositional tuning data is critical in compositional generalization.

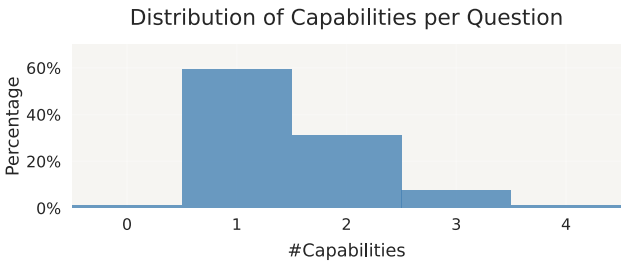


Figure 5. **Distribution of Compositional Complexities in LLaVA-665K samples.** Majority of questions (59.2%) use one atomic capability, followed by 30.9% using two.

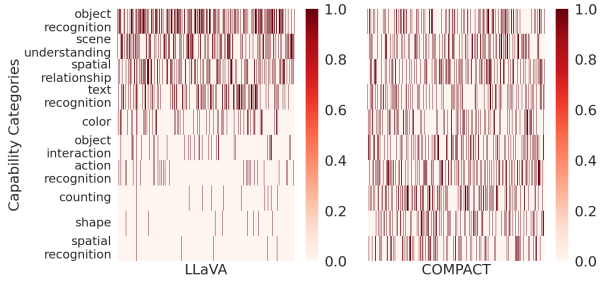


Figure 6. **Comparison of Capability Distribution.** The heatmaps show the frequency of each atomic capability in LLaVA (*left*) and COMPACT (*right*) samples. The capabilities are sorted by frequency based on the LLaVA capability distribution, with more common capabilities appearing closer to the top. In LLaVA, the distribution is notably imbalanced: object recognition and scene understanding are some of the most frequent, while shape and spatial recognition are less prevalent. In contrast, COMPACT exhibits a more balanced distribution across all capability categories.

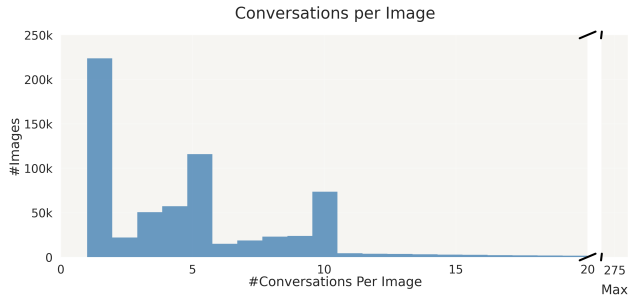


Figure 7. **Distribution of conversations per image in LLaVA-665K.** The overwhelming majority of images (97.69%) have $\leq$20 conversation pairs. The average of number of conversations per image is 5.18 ($\sigma = 5.62$). A small subset (2.31%) exceeds 20 conversations, which includes a sample with the maximum length of 275. Total conversations: 3,444,246.

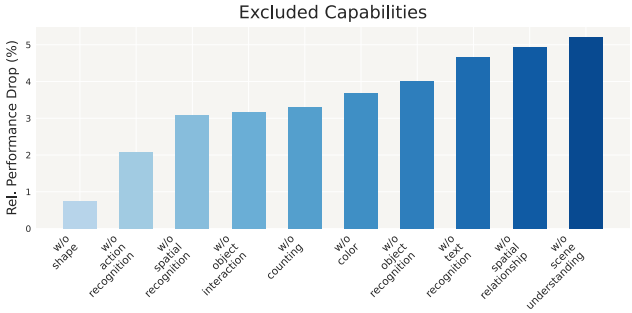


Figure 8. **Leave-One-Out Analysis on Atomic Capabilities.** We measure the average performance degradation across benchmarks by excluding an atomic capability from training. Higher drop indicates higher importance of the atomic capability. Excluding scene understanding and spatial relationships have the largest impact, while that of excluding shape and action recognition are modest.

Impact of Atomic Capability Coverage. To validate our choice of atomic capabilities and understand their relative importance, we conduct a leave-one-out analysis by systematically excluding questions that require a specific capability while keeping the total number of training examples fixed. As shown in Fig. 8, scene understanding and spatial relationship emerge as the most critical capabilities, with each of their exclusion leading to a significant performance drop (5.22% and 4.93% respectively). Text recognition and object recognition are also essential (4.65% and 4.03% drops). The exclusion of capabilities like shape attribution and action recognition have a smaller impact (0.74% and 2. 08%). This analysis validates our selection of atomic capabilities by demonstrating that each capability contributes meaningfully to overall performance without being redundant.

Effect of Compositional Complexity Range. To isolate the effect of the range of compositional complexities while controlling for data quality, we generate three sets of 16K-

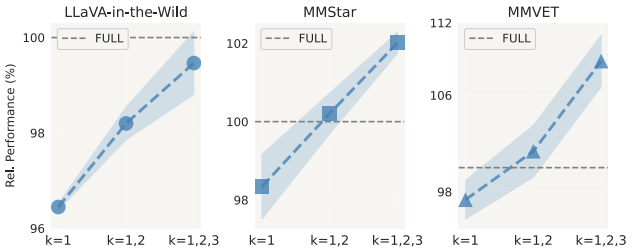


Figure 9. **Compositional Complexity Analysis:** Performance comparison of models trained with different compositional complexities. $k = 1$ refers to only one atomic capability per question, $k = 1, 2$ to both single and dual capabilities, and $k = 1, 2, 3$ to single, dual, and triple capabilities. Results show consistent improvements as the range of compositional complexities increases.

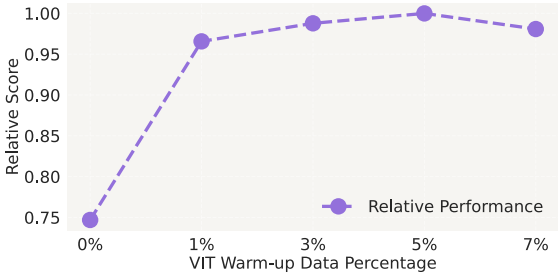


Figure 10. **Impact of Instruction Tuning Data Ratio on Performance.** Relative performance of models trained on COMPACT mixed with different ratios instruction tuning data from LLaVA-665K [24]. The x-axis is the percentage of LLaVA-665K [24] used as instruction tuning data, and the y-axis is the average relative score across benchmarks. The performance improves significantly with a small percentage of instruction tuning data and stabilizes around 5%.

sample compositional tuning data, each with $k = 1$, $k = 1, 2$ or $k = 1, 2, 3$, using identical Gemini-2.0-Flash [36] configurations. For fair comparison, we maintain consistent sample counts and use an identical set of images in all three settings. The model trained on only $k = 1$ (single capability per question) underperforms the model trained on $k = 1, 2, 3$ compositional tuning data on multi-capability benchmarks: MM-Vet [43] (28.82 vs. 29.22), LLaVA-W [24] (66.1 vs. 68.5) and MMStar [6] (34.53 vs. 35.11). This shows that although the model trained on $k = 1$ data can solve tasks with lower compositional complexity, it struggles to perform in higher compositional complexities.

As shown in Fig. 9, increasing the range of compositional complexities leads to consistent improvements on all three benchmarks. Training on $k = 1, 2, 3$ compositional tuning data achieves the highest performance on MM-Vet [43] (32.61) and MMStar [6] (0.3577), demonstrating that exposure to more complex compositional patterns during training enhances the model's ability to handle complex multi-capability tasks. Surprisingly, the model achieves 112% performance on MM-Vet [43] with only 16k compositional tuning data compared to the LLaVA-665K [24] baseline, suggesting that a balanced mixture of different compositional complexities improves data efficiency.

Impact of Instruction Tuning Data Ratio. We vary the amount of instruction tuning data sampled from the LLaVA-665K [24] VIT data to understand the impact of the mixing ratio on model performance. In order to isolate the effect on visual instruction following, we exclude $k = 0$ conversations (approximately 1.1% of the questions), which have minimal relevance to visual capabilities. As we scale the VIT subset from 0% (pure compositional tuning) to 7% of LLaVA-665K [24], we observe an upward trend in performance, indicating that the role of the instruction tuning data is crucial. Fig. 10 shows that without instruction tuning data (0%), COMPACT achieves only 74.69% of the performance relative to LLaVA-665K [24] VIT. Increasing the instruction tuning data to just 1% of the VIT data significantly improves the relative performance to 96.56%. However, further scaling gives diminishing returns, with 3% reaching 98.77% and 5% achieving nearly identical relative performance (99.99%). Interestingly, taking 7% from the VIT data causes a slight decrease to 98.07%, indicating that 5% represents an optimal balance between instruction tuning and compositional tuning data in terms of data efficiency and performance. These results suggest that instruction following capability is potentially orthogonal to the capabilities of the base model and the atomic visual capabilities, and can be acquired with minimal instruction tuning data.

F. Additional Discussions

Limitations. Our approach faces two key limitations. First, we rely on data generated from closed-source models (i.e., Gemini), which potentially introduce their compositional limitations and biases to our dataset. Additionally, this data generation process is costly, which could pose challenges for reproducibility. To support future research, we will publicly release the data generated in this project. Second, our approach focuses on the compositionality of vision-centric capabilities. Therefore, our approach may not be optimal for addressing knowledge-intensive tasks that lie outside the scope of visual reasoning. See Appendix §**??** for a detailed discussion on knowledge-intensive task results.

Future Work. We aim to extend COMPACT to accommodate higher-order compositional complexity (k $>$ 3). Currently, our data recipe only generates data up to k $=$ 3 due to the decreasing reliability of closed-source models at higher compositional complexities. Specifically, as the number of atomic capabilities increases, their integration tends to be more inconsistent, ambiguous, or erroneous. Future work could explore hierarchical composition approaches or hybrid data generation pipelines that

combine multiple sources and verification steps to improve performance on higher compositional complexities. Additionally, experimenting with explicit reasoning approaches (e.g., step-by-step decomposition [33]) could further improve the model's ability to solve complex tasks while retaining data efficiency.

G. Visualizations

Qualitative Comparison. We provide qualitative visualizations that compare the outputs from our compositionally-tuned COMPACT model and the LLaVA-665K VIT model. Examples in Fig. 11 highlight the importance of compositional tuning for handling complex multi-capability tasks ($k \geq 3$). These cases demonstrate COMPACT model's enhanced ability to integrate multiple visual capabilities, while showing the baseline model's difficulty with such compositionally complex queries.

Zero-Capability Samples in LLaVA-665K. We identify a subset of samples in the LLaVA-665K dataset that require no visual capabilities, which we refer to as zero-capability samples. These include general knowledge queries, subjective prompts, or requests that can be answered without inspecting the image at all. While such data may still be useful for instruction following, it does not contribute to the development of vision-centric skills. In our analysis, we find that approximately 1.1% of the questions in LLaVA-665K fall into this zero-capability category.

Zero-Capability Samples

Zero-Capability Questions:

- How is the weather?
- Should I move to London?
- Can you provide some information about the Emirates airline?
- Give me a long list of what duties are considered rental activity
- Have the cat declare her new name as ruler
- rewrite it from the perspective of an expensive therapist
- Can you tell me how to prepare a Colombian dish
- how to do coding
- Can you explain Map Reduce to me?
- A 35 year old patient presented to the emergency department with shortness of breath. Before this, he was at a crowded event. He does not have a history of diabetes or high blood pressure. He had a positive PCR test at an outside hospital. What should be the next steps for the physician?
- please convert those snomed codes to FHIR
- I'm running a used car dealership, what are some emerging opportunities for me brought by large language models like GPT-3?
- answer it again in Chinese
- you are a legislator. You are asked to come up with a framework for new legislation that adances the science of reading for grades K-3. Write that model legislation.
- I'm looking to create a podcast, can you help me?

COMPACT Data Visualization. We provide a visualization of the COMPACT dataset to provide insights into its compositional structure. Figs. 12 and 13 show selected examples from the COMPACT dataset. Each question is generated from a combination of k atomic capabilities. These cases demonstrate our model's enhanced ability to integrate multiple visual capabilities simultaneously, while the baseline model often struggles with such compositionally complex queries.

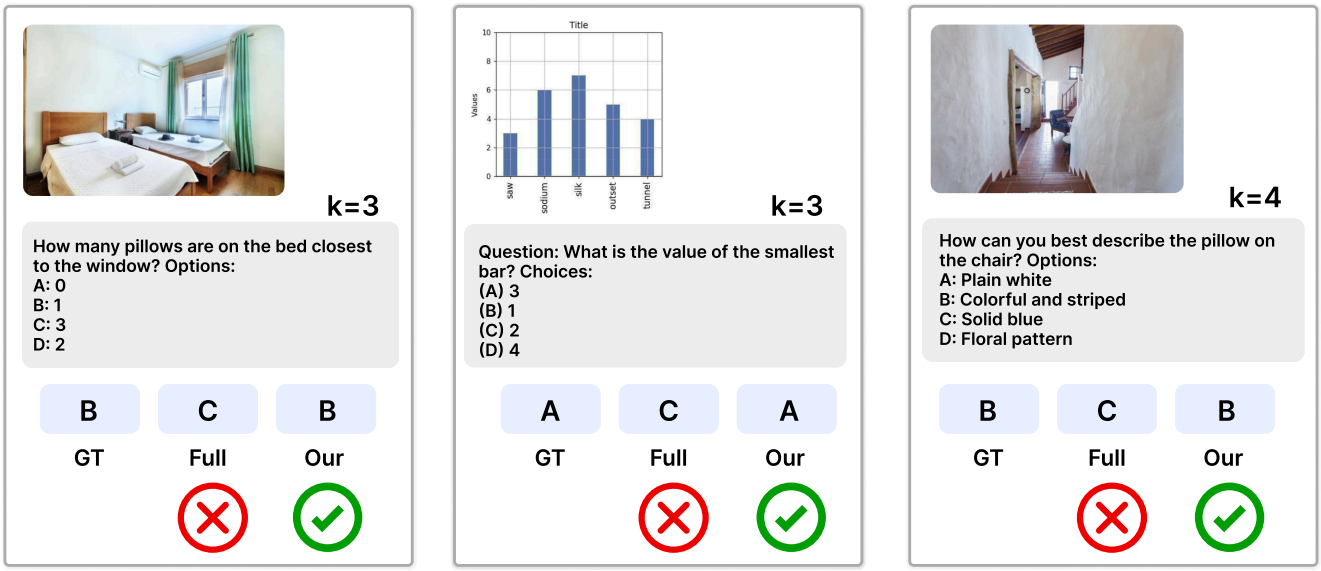


Figure 11. **Qualitative comparison of model outputs.** Examples showing responses from our compositionally-tuned COMPACT model and LLaVA-665K [24] VIT model on complex queries that require multiple capabilities ($k \geq 3$). Our model demonstrates better integration of visual capabilities which leads to more accurate responses.

How many people are visible?

k=1, counting

Two

What color is the vest worn by the person interacting with the train?

k=2,
color,
object_interaction

Orange

What color are the rectangular train cars next to the man in uniform?

k=3,
color,
shape,
spatial_relationship

Maroon

How many laptops are on the table?

k=1, counting

Two

What object is reflecting in the screen of the left laptop?

k=2,
object_recognition,
object_interaction

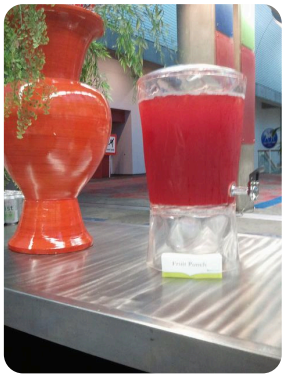
Camera

How many laptops displaying the text 'Eee PC' are visible in this indoor scene?

k=3,
scene_understanding,
counting,
text_recognition

Two

Figure 12. **Visualization of COMPACT Compositional Tuning Samples.**



What is next positioned next to the punch dispenser?

k=1, spatial_relationship

Vase

What color liquid is being dispensed from the beverage dispenser?

k=2,
action_recognition,
color

Red

What action is depicted on the sign behind the vase?

k=3,
action_recognition,
spatial_recognition,
spatial_relationship

Hockey

What time of day does the lighting suggest?

k=1, spatial_recognition

Sunset

How many green motorcycles are visible?

k=2,
color,
counting

One

What color is the motorcycle parked on the city street?

k=3,
color,
object_recognition,
scene_understanding

Red

Figure 13. **Visualization of COMPACT Compositional Tuning Samples.**