# 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 [31, 39], and how the model responds to a prompt can be critical to the evaluation of its performance [9]. The misalignment between the model response and the format demanded by question can cloud precise evaluation of its capabilities [1, 10, 26]. Visual instruction tuning [18, 20] is proposed to adapt MLLMs to respond appropriately to diverse question formats (e.g., multiple-choice, short- and long-response questions). 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 [12], recent work [8] has shown that optimizing for response formatting potentially limits the response quality of language models.

Unlike VIT [20], which focuses on response formatting and simple capabilities of MLLMs, our approach explicitly models compositionality, improving performance on tasks that are more complex in capability space. Our method leverages the response formatting potential of VIT and complex reasoning potential of compositionally generated synthetic conversations to create a more optimal data recipe.

**Compositionality in LLMs and MLLMs.** Compositionality is the ability to combine simpler concepts or elements into more complex expressions [6]. In the context of visual reasoning in MLLMs, compositional capability refers to a model's ability to perform complex tasks by combining multiple capabilities [11], 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 [35, 38], but generalizations to the realms of MLLMs have been largely incomplete. Some studies highlight that while MLLMs do show signs of compositional capability [24], they struggle when constituting components and their combined patterns are not strongly learned or missing during training [2]. Furthermore, previous works have focused on limited domains such as geometry [3], visual recognition, and language [5] while others employ a relaxed definition of compositionality as a sequential array of tasks rather than fundamentally combining them.

However, studies show that general visual reasoning requires strong compositional ability [37]. In order to train compositional capabilities in MLLMs, modeling compositionality in the training data is necessary. Our method leverages these findings to create a data recipe for tackling complex visual reasoning tasks across wider domains.

**Data Efficiency in MLLMs.** VIT is a data and compute-heavy step in training [34]. Studies have found that the performance of MLLMs can be reproduced with better techniques, suggesting that the amount of data needed for VIT can be reduced. For example, recent work has developed effective VIT data recipes leveraging data selection methods and curating higher-quality training datasets [14, 22]. ICONS [32] 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 reasoning capabilities even further [16].

However, these methods treat compositionality as a byproduct of scale rather than as a learnable capability. COMPACT formalizes atomic capabilities and systematically incorporates their combinations into the 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. 3 shows the taxonomy of atomic capabilities and their detailed descriptions.

## C. Additional Experiment Details

**Benchmarks.** We evaluate the trained model on established multimodal benchmarks with multi-capability settings: 1) **MM-Vet** [36] includes 16 types of complex multimodal tasks integrated from 6 core capabilities (recognition, OCR, knowledge, language generation, spatial awareness, and math). 2) **MME** [7] is a manually designed dataset for MLLM evaluation which contains 14 perception and cognition related subtasks. 3) **LLaVA-in-the-Wild** [20] is an open-ended visual question answering benchmark that asks complex reasoning questions on real-world images. 4) **SEED-Bench-2-Plus** [15] evaluates text-rich visual comprehension skills of MLLM that focuses on charts, maps, and webs. 5) **MMStar** [4] contains human reviewed question-answers from existing benchmarks that span 6

Table 3. **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? |

core capabilities (fine-grained perception, coarse perception, mathematics, science & technology, logical reasoning and instance reasoning). 6) **CV-Bench** [30] is a MLLM benchmark specialized for 2D and 3D visual understanding which includes spatial relationship, object count, relative distance, and depth order. 7) **TextVQA** [28] evaluates visual reasoning capabilities related to texts in the image. 8) **InfoVQA** [23] measures reasoning and arithmetic skills using infographic images.

We provide a system prompt for capability analysis that identifies the required capabilities for a given question. We also provide implementation details for the compositional question generation and verification systems used in our experiments. The question generation system prompt generates multi-capability questions through structured guidelines that blend different capabilities naturally and can only be answered by checking the images. The question verification system prompt checks if the questions meet all requirements and include the required number of capabilities, without relying on subjective interpretations or having compositional flaws. We provide the system prompt below.

**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"]

## D. Additional Analysis

**Visual Capability Distribution in LLaVA-665K.** Fig. 4 shows the distribution of the number of capabilities required per question. We sampled 1,000 data points from LLaVA-665K which contains 5,668 questions, and analyzed their compositional complexity using Gemini-2.0-Flash. The mean compositional complexity of the questions is approximately 1.5, and mode is 1, with 59.2% of questions utilizing only one capability. An additional 30.9% of questions use 2 capabilities, resulting in 90.1% of questions requiring 2 or less atomic capabilities. Only 0.1% of questions require 5 capabilities, and a small fraction (0.2%) require as many as 10 capabilities (e.g., "Question: Describe this photo in detail."). LLaVA-665K VIT's training paradigm exhibits a *complexity cliff*—a steep decline in downstream task performance as $k$ increases (Tab. 2). This performance gap between different compositional complexities shows how VIT's complexity-agnostic training leaves models unpre-

**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)

pared for tasks requiring compositional generalization. Interestingly, we observe that 1.1% of the data includes questions with zero capabilities, as illustrated in Fig. 4. Fig. 5 shows the frequency of capabilities in the LLaVA-665K dataset, with object recognition (38.97%) and scene understanding (28.58%) being 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%).

COMPACT preserves sampled image content from the LLaVA-665K VIT dataset [20] when constructing new conversations with compositional structure. This enables a fair comparison between existing methods' and COMPACT's ability to extract richer and more structured information from images. We further adjust the ratio between our compositional data and the original VIT subset to study

**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

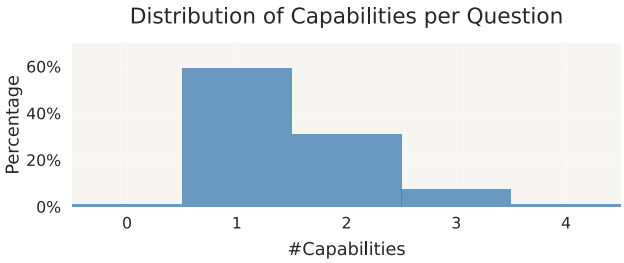


Figure 4. **Distribution of Compositional Complexities in LLaVA-665K samples.** Majority of questions (59.2%) use one atomic capability, followed by 30.9% using two.

how different mixtures affect the performance (§E). This allows us to find the optimal balance between the preservation of instruction-following capability and the development of compositional capabilities.

**Analysis of Conversation Distribution in LLaVA-665K.** Fig. 6 shows the distribution of conversations per image in LLaVA-665K, where 97.69% of images contain $\leq$20 conversation pairs. The average of 5.18 conversations per image ($\sigma = 5.62$) reveals that the data is heavily concentrated around shorter interactions. This concentration of data, where over 93.6% of samples fall below the 10-pair threshold, directly guided our decision to anchor the number of conversations in the compositional tuning dataset to this range. By aligning with the dominant distribution, we ensure a fair comparison.

**Analysis of Limited Performance Gains on Knowledge-Intensive Tasks.** While our compositional tuning approach shows general improvements across various benchmarks, we observe more modest gains on knowledge-intensive tasks. Table 4 shows the performance comparison across different model variants on OK-VQA, MMMU,

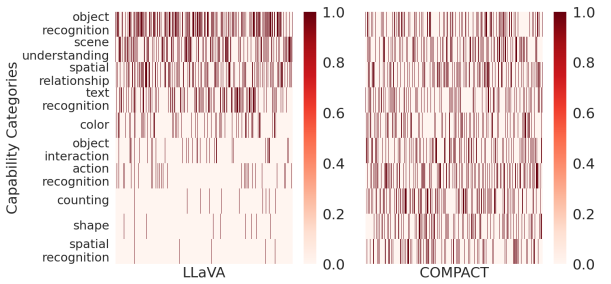


Figure 5. **Comparison of Capability Distribution.** The heatmaps show the frequency of each atomic capability in LLaVA (*left*) and COMPACT (*right*) samples. The capabilities are sorted by frequency based on the LLaVA capability distribution, with more common capabilities appearing closer to the top. 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, our COMPACT exhibits a more balanced distribution across all capability categories.

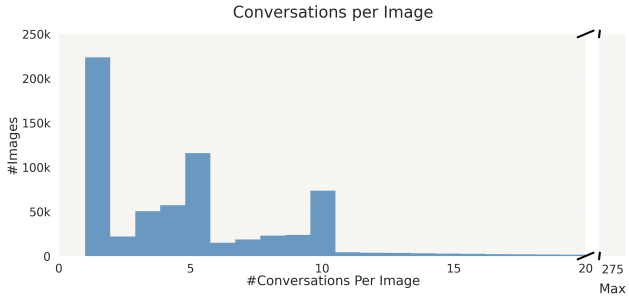


Figure 6. **Distribution of conversations per image in LLaVA-665K.** The majority of images (97.69%) have $\leq$20 conversation pairs, with an average of 5.18 conversations per image ($\sigma = 5.62$). A small subset (2.31%) exceeds 20 conversations, including a sample with a maximum length of 275. Total conversations: 3,444,246.

and MMMU-Pro benchmarks. With our 32k sample training set, we see relatively small improvements over the baseline: OK-VQA (50.02% vs 49.30%), MMMU (33.89% vs 32.89%), and MMMU-Pro (20.23% vs 18.15% for standard tasks, 11.91% vs 11.44% for vision tasks). Notably, even when training on the full dataset, the improvements on MMMU remain limited (33.89%). While knowledge-related tasks are not our main focus, this inspires future work to design compositional tuning approaches covering broader capabilities, which is outside of our vision-centric capability scope.

## E. Ablations

We conduct a series of ablation studies to investigate key design considerations (atomic capability coverage, compo-

Table 4. **Limited Performance on Knowledge-Intensive Benchmarks.** Comparison shows modest improvements over baseline on tasks requiring 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.) |
| Baseline | 49.30 | 32.89 | 18.15 | 11.44 | 92.0% |
| Ours (32k) | 50.02 | 33.89 | 20.23 | 11.91 | 96.6% |
| Full | 57.96 | 33.89 | 20.12 | 11.97 | 100% |

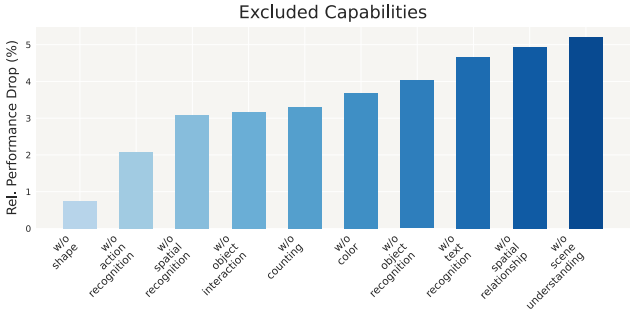


Figure 7. **Leave-One-Out Analysis on Atomic Capabilities.** We measure the average performance degradation across benchmarks by excluding an atomic capability from training. Higher drop indicates higher importance of the atomic capability. Excluding scene understanding and spatial relationships have the largest impact, while that of excluding shape and action recognition are modest.

sitional complexity, and instruction tuning ratio) in COMPACT. Unless otherwise specified, all experiments use 16K $k = 1, 2, 3$ compositional tuning data.

**Impact of Atomic Capability Coverage.** To validate our choice of basis capabilities and understand their relative importance, we conduct a leave-one-out analysis by systematically excluding questions requiring specific capabilities. As shown in Fig. 7, scene understanding and spatial relationship emerge as the most critical capabilities, with their exclusion leading to significant performance drops. Text recognition and object recognition are also essential, while excluding more specialized capabilities like shape attribution and action recognition has a smaller impact. 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.** To isolate the impact of compositional complexity while controlling for data quality, we generate $k = 1$, $k = 1, 2$ and $k = 1, 2, 3$ compositional tuning data using identical Gemini-2.0-Flash configurations. For fair comparison, we maintain consistent sample counts and use identical images across all three settings. The model trained on only $k = 1$ data (single capability per question) underperforms the model trained on $k = 1, 2, 3$ compositional tuning data on complex reasoning benchmarks. This shows that although the model trained on $k = 1$ data can solve tasks with lower compositional complexity, it struggles with generalization to those with higher compositional complexity.

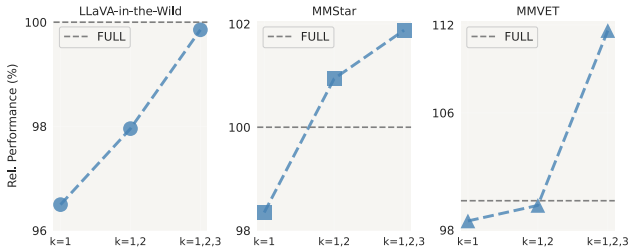


Figure 8. **Compositional Complexity Analysis:** Performance comparison of models trained with different compositional complexities. $k = 1$ means the model is only trained on a single atomic capability per question, $k = 1, 2$ to both single and dual capabilities, and $k = 1, 2, 3$ to single, dual, and triple capabilities. Results show consistent improvements as compositional complexity increases.

As shown in Fig. 8, increasing the compositional complexity of the training data leads to consistent improvements on all three benchmarks. Training on $k = 1, 2, 3$ compositional tuning data achieves the highest performance on MM-Vet (32.61) and MMStar (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 with only 16k data points compared to the LLaVA-665K baseline, suggesting that a balanced mixture of different compositional complexities improves data efficiency.

**Different Instruction Tuning Data Ratio.** We vary the amount of instruction tuning data sampled from the LLaVA-665K VIT dataset to understand its impact on model performance. Scaling this subset from 0% (pure compositional tuning) to 7% of LLaVA-665K, we observed an upward trend in performance which indicates that the role of instruction tuning data is crucial. As shown in Fig. 9, without instruction tuning data (0%), the model achieves only 74.69% of the performance relative to the model trained on LLaVA-665K, as it struggles to follow instructions correctly. Increasing the instruction tuning data to just 1% of LLaVA-665K significantly improves the relative performance to 96.56%. However, further increases show diminishing returns, with 3% reaching 98.77% and 5% achieving nearly identical performance (99.99%) relative to the model trained on LLaVA-665K. Interestingly, taking 7% of LLaVA-665K causes a slight decrease to 98.07%, suggesting that 5% represents an optimal balance

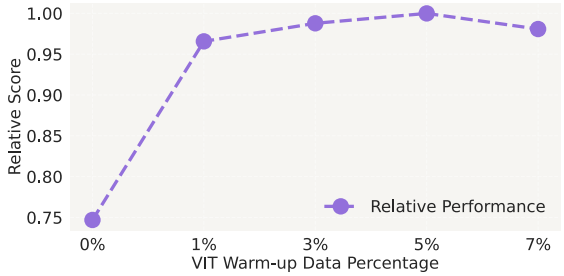


Figure 9. **Impact of Instruction Tuning Data Ratio on Performance.** Relative performance of models trained on COMPACT with different ratios of instruction tuning data compared to standard VIT model trained on LLaVA-665K. The x-axis is the percentage of LLaVA-665K used for instruction tuning data, and the y-axis shows the average relative score across benchmarks. The performance improves significantly with a small percentage of instruction tuning and stabilizes around 5%, indicating efficient learning with reduced data.

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. Second, our approach focuses on compositionality of visual capabilities. Therefore, our approach may not be optimal for addressing knowledge-intensive tasks. See Appendix D 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 method 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) could further improve the model's ability to solve complex tasks while retaining data efficiency.

## G. Visualizations

**Qualitative Comparison.** We provide qualitative visualizations comparing outputs from our COMPACT compositionally-tuned model against the model trained with LLaVA-665K. Fig. 10 shows selected examples that highlight how compositional tuning improves handling of complex multi-capability tasks ($k \geq 3$). 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.

**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

**COMPACT Data Visualization.** We provide a visualization of the COMPACT dataset to provide insights into its compositional structure. Fig. 11 shows selected examples from 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.

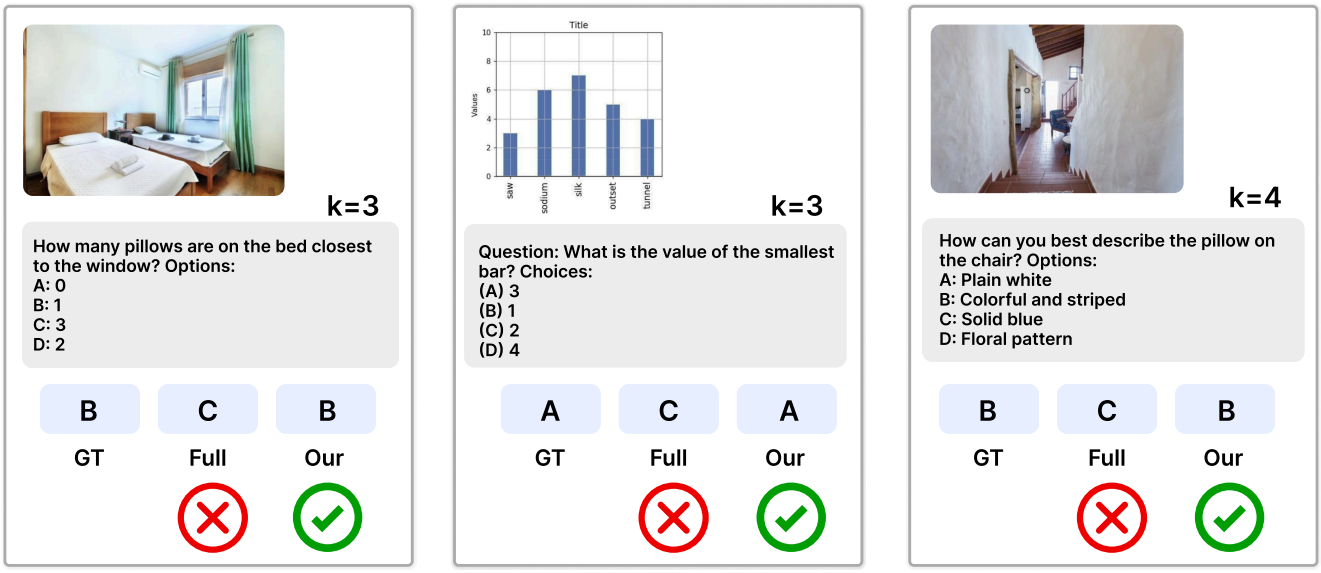


Figure 10. **Qualitative comparison of model outputs.** Examples showing responses from our compositionally-tuned model versus the model trained with LLaVA-665K on complex queries requiring multiple capabilities ($k \geq 3$). Our model demonstrates better integration of multiple visual reasoning capabilities, leading to more accurate and comprehensive responses.

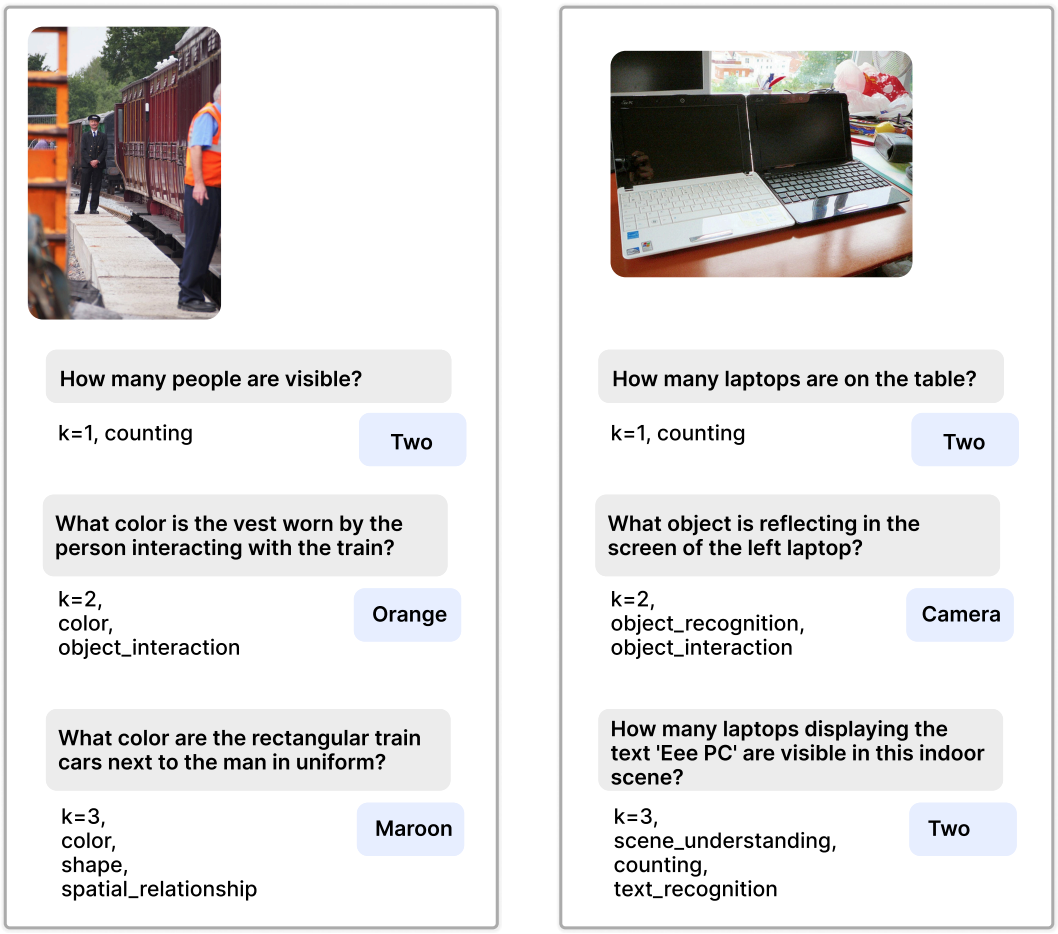


Figure 11. **Visualization of COMPACT Compositional Tuning Samples.** Here we provided two examples from our COMPACT compositional tuning dataset, including $k = 1, 2, 3$.