
Unveiling the Compositional Ability Gap in Vision-Language Reasoning Model

Anonymous Author(s)

Affiliation

Address

email

1 Limitations

2 While our study provides valuable insights into the compositional generalization of vision-language
3 models, it also has several limitations that point to important directions for future research. First, our
4 benchmark focuses on synthetic visual reasoning tasks (e.g., shape area, spatial position, and their
5 composition), which are well-controlled but may not fully capture the complexity and ambiguity
6 of real-world multimodal scenarios. Although synthetic data allows precise supervision and clear
7 reasoning traces, the domain shift to natural images remains unaddressed in this work. Second, the
8 diagnostic tasks in ComPABench are primarily designed around two types of compositional reasoning
9 (geometric and spatial). While this allows a focused analysis, it does not cover other important
10 types of reasoning such as causal, or commonsense multimodal reasoning, which could affect model
11 behavior in broader contexts. Lastly, while RL-Ground demonstrates strong gains, its reliance on
12 structured captions and progress rewards assumes access to intermediate signals. Applying the same
13 method to open-ended real-world tasks may be less straightforward and require new mechanisms for
14 unsupervised or weakly-supervised reward shaping. We hope future work will explore scaling our
15 findings to more realistic multimodal benchmarks, improving robustness across diverse task types,
16 and relaxing the reliance on fine-grained supervision signals.

17 Broader Impacts

18 This work contributes to the understanding and advancement of compositional reasoning in VLMs.
19 By introducing a diagnostic benchmark and evaluating different post-training strategies, we aim to
20 improve the transparency and robustness of VLMs in performing multimodal reasoning. Improving
21 compositionality in VLMs can benefit applications that require complex reasoning over visual inputs,
22 such as educational tools, scientific assistants, and visual question answering systems in safety-critical
23 domains (e.g., medical or industrial analysis). Our findings also highlight potential failure modes
24 of current models in integrating visual and symbolic knowledge, offering actionable insights for
25 designing more interpretable and trustworthy systems. However, as with any progress in general-
26 purpose AI capabilities, there are risks of misuse. Enhanced reasoning abilities could be exploited
27 to generate more persuasive misinformation grounded in visual content or to automate decisions in
28 sensitive contexts without sufficient human oversight. We emphasize that our benchmark is designed
29 for diagnostic and research purposes, and we encourage responsible use of our findings and dataset.

30 A Benchmark Details

31 **Dataset Curation.** To facilitate the evaluation of multimodal compositional reasoning abilities in
32 VLMs, we construct three diagnostic benchmark tasks: Shape Area, Grid Position, and Area-Position
33 Composition as shown in Table 1. Each dataset consists of synthetic images paired with natural
34 language questions, thinking path for supervised finetuning, and final answers.

Property	Shape Area Task	Grid Position Task	Area-Position Composition
Total Samples	4K(train), 500(eval)	4K(train), 500(eval)	500(eval)
Image Resolution	512×512	512×512	up to 1000×1000
Grid Size	N/A	3×3 to 10×10	3×3 to 10×10
Shapes per Image	2 to 6	2 to 6	2 to 6
Color Palette	10 predefined (unique per image)	same	same
Answer Format	Integer (area)	Grid position (e.g., (2, 4))	Integer (area)
Rationale Trace	LaTeX-style formula	Stepwise distance trace	Distance + Area reasoning

Table 1: Dataset statistics and task features for Shape Area, Grid Position, and Area-Spatial Compositional reasoning benchmarks.

The **Shape Area Task** involves computing the area of a queried geometric shape from an image containing 2 to 6 shapes with overlaid dimension labels. Shapes are selected from *square*, *rectangle*, *right triangle*, and *trapezoid*, and each is assigned a unique color from a fixed palette of 10. The thinking path is provided as a LaTeX-style symbolic formula, and the final answer is a rounded integer.

In the **Grid Position Task**, the model must identify the grid index of the shape closest to a target using Manhattan distance on a 3×3 to 10×10 grid. Each image contains 2 to 6 non-overlapping shapes positioned within discrete grid cells. Shapes are rendered with unique colors and grid-aligned placement to maintain spatial consistency. The thinking path includes the identity of the target, distances to other shapes, and a final reasoning step determining the closest shape.

The **Area-Position Composition Task** evaluates compositional reasoning by requiring the model to compute the combined area of the target shape and its nearest neighbor. This task fuses the geometric and spatial reasoning elements of the previous two tasks. The image generation process follows the same principles, and the reasoning step combines both the spatial distance trace and symbolic area calculations.

All datasets are rendered at high resolution (512×512 or higher). We generate 4K training samples and 500 evaluation samples for the individual tasks, and 500 evaluation samples for the compositional setting.

Pure-Text Variant. To assess the modality gap in compositional reasoning, we also construct pure-text counterparts for the above three tasks. Instead of image inputs, the pure-text version presents the shape attributes (e.g., type, color, and dimensions or grid positions) directly in natural language. Each sample includes a textual description of the visual scene and a question that matches the reasoning objective of its multimodal counterpart. The same symbolic thinking steps and answer format are retained. This setup enables controlled comparisons across input modalities, isolating the effects of visual grounding.

Out-of-Distribution (OOD) Evaluation. To evaluate generalization beyond seen task objectives, we curate OOD variants for each of the three benchmarks. For the **Shape Area Task**, the question changes from computing the total area to identifying the largest-area shape. For the **Grid Position Task**, the query asks for the shape *farthest* from the target in Manhattan distance instead of the nearest. In the **Area-Position Composition Task**, the model is asked to identify the larger of the target and farthest shape, and then report its area. Each of these OOD variants consists of 500 evaluation-only samples. This setting probes the model’s robustness to distributional shifts in task semantics, while keeping the input format consistent.

B Ablation Study on RL-Ground

To understand the individual contributions of RL-Ground’s components, we conduct an ablation study using the Qwen2.5-VL-Instruct-7B model on three evaluation tasks: Shape Area, Grid Position, and Compositional Reasoning. As shown in Table 2, we compare the base RL setup against two variants—adding either the <caption> format or progress reward supervision—and evaluate the full RL-Ground configuration that combines both.

Method	Shape Area	Grid Position	Compositional
RL (baseline)	74.6	83.2	31.2
RL + Caption Format	87.4	84.6	28.0
RL + Progress Reward	76.0	85.8	39.6
RL-Ground	73.8	88.4	52.8

Table 2: Ablation results for RL-Ground components on Qwen2.5-VL-Instruct-7B.

74 Adding the **caption format alone** improves performance on Shape Area (87.4% vs. 74.6%) and Grid
75 Position (84.6% vs. 83.2%), likely due to enhanced perceptual grounding from explicitly verbalizing
76 the visual scene. However, it slightly underperforms on the compositional task, suggesting that format
77 change alone may not encourage compositional multi-step reasoning.

78 Adding the **progress reward** improves compositional accuracy substantially (39.6% vs. 31.2%) and
79 moderately boosts performance in Grid Position. This indicates that dense intermediate supervision
80 helps the model build more robust reasoning chains.

81 Finally, **RL-Ground**, which integrates both mechanisms, achieves the best overall compositional
82 performance (52.8%) and the highest Grid Position score (88.4%), despite drop in Shape Area. These
83 results validate that combining image-to-text conversion with progress-based reward creates strong
84 advantages for compositional generalization.

85 C Code & Dataset

86 To ensure the reproducibility of our experiments, we submit both the full codebase and the multimodal
87 datasets used in our training.

88 **Dataset Submission.** We provide the complete 8K multimodal training samples used for the Shape
89 Area and Grid Position tasks, comprising 4K training samples per task. Each sample consists of a
90 rendered image, a natural language question, symbolic reasoning steps, and a final answer. These
91 datasets serve as the foundation of our proposed **CompABench** benchmark and are constructed to
92 test compositional reasoning through both geometric and spatial tasks. All dataset generation scripts
93 are included under:

94 `ComPA/src/r1-v/local_scripts/`

95 This directory contains tools for rendering geometric shapes, generating grid layouts, and annotating
96 reasoning traces.

97 **Training and Evaluation.** To train or evaluate models, users should convert their data to match our
98 provided JSONL format. Specifically:

- 99 • Training data should be formatted in the same structure as the submitted shape area and grid
100 position files.
- 101 • Evaluation prompts should follow the format specified in:

102 `ComPA/src/eval/prompts/spatial-reasoning-eval-new.jsonl`

103 We include complete scripts for:

- 104 • Supervised fine-tuning (SFT)
- 105 • Reinforcement learning with GRPO (RL)
- 106 • Our proposed **RL-Ground** framework

107 Setup instructions, including environment configuration and training commands, are described in the
108 accompanying README.md. By providing both the datasets and training pipeline, we aim to enable
109 rigorous and reproducible evaluation of multimodal compositional reasoning models.