GroundingME: Exposing the Visual Grounding Gap in MLLMs through Multi-Dimensional Evaluation

Supplementary Material

A. Detailed Results

A.1. Subtask Results in Analysis

We report the detailed subtask results for the L-1 categories omitted from the analysis section. Tab. 1 presents the detailed subtask performance of models when enabling thinking mode, supplementing the analysis on performance gain discussed in §5.1. Tab. 2 shows the out-of-domain subtask performance of the fine-tuned Qwen3-VL-8B-Instruct model, complementing the overall results presented in §5.3.

Table 1. Subtask performance of different models by enabling thinking mode.

Model	Total	Dis.	Spa.	Lim.	Rej.
Qwen3-VL-8B	34.3	52.5	43.0	33.3	4.5
Qwen3-VL-32B	46.9	65.7	70.0	36.0	9.5
Qwen3-VL-A3B	39.2	53.4	53.3	38.0	5.5
Qwen3-VL-A22B	49.8	65.2	73.7	45.0	5.5
GLM-4.5V	34.0	52.5	45.3	30.3	4.0
MiMo-VL-7B-RL	24.1	46.6	28.7	17.0	5.0
Seed-1.6-V.-250815	46.5	59.3	72.7	41.7	1.5

Table 2. Subtask performance of fine-tuned Qwen3-VL-8B-Instruct under different SFT data ratios.

Neg.:Pos.	Total	Dis.	Spa.	Lim.	Rej.
1:8	27.0	57.4	24.7	24.3	3.5
1:4	25.0	49.5	25.3	19.3	8.0
1:2	28.7	54.4	28.3	23.0	11.4
1:1	28.5	46.6	26.3	26.3	16.4
2:1	26.0	40.2	24.0	17.0	27.9

A.2. Main Results across Various IoU

For the evaluation presented in the main results table, we further report the accuracy of all models on the entire GroundingME and three L-1 categories (the Rejection category is excluded, as its accuracy is independent of the IoU threshold) across different IoU thresholds in Tab. 3. New metrics include Accuracy@0.75, Accuracy@0.9, and mAcc. The mAcc is defined as the mean accuracy calculated over the range of IoU thresholds [0.5, 0.95], sampled at intervals of 0.05.

B. Cross-Dimensional Analysis

B.1. Cross-Dimensional Imbalance.

Fig. 1 illustrates a significant performance imbalance across dimensions. A consistent performance hierarchy emerges across the four L1 types: *Dis.*>*Spa.*≈ *Lim.*>*Rej.* Similar inconsistencies are observed at L2 types (e.g., *Sta.* ranks lowest within *Dis.*, and *Cnt.* lower than *Rel.* within *Spa.*). These results highlight a systematic bias in current MLLMs.

B.2. Subtask Ranking Difference.

Model rankings exhibit significant divergence across different dimensions. In Fig. 1, Seed-1.6-V. (#2) ranks only #7 in the *Dis. Txt.* task, suggesting a potential weakness in OCR capabilities, yet it achieves #1 in all *Spatial* tasks. Such discrepancies reveal potential imbalances in training data, providing a clear direction for domain-specific enhancements.

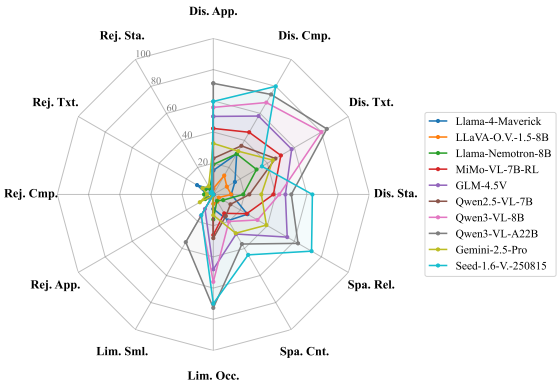


Figure 1. Cross-Dimensional Performance.

B.3. Model Family Behaviors.

Fig. 2 explores the behaviors of models within the Qwen3-VL dense model family (2B-32B). We observe: (1) Models of all size score 0 on *Rejection*, demonstrating high model-family consistency. (2) Performance on *Discriminative* and *Spatial* correlates positively with model scaling, while no obvious trend on *Limited*. (3) Subtask decomposition reveals the anomaly of *Limited* stems from *Small*, providing insight for further analysis.

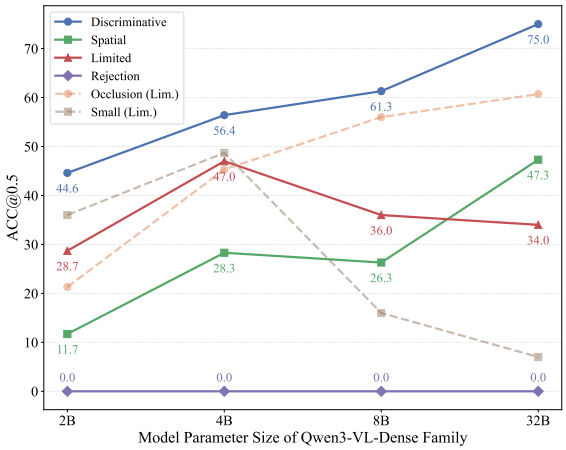


Figure 2. Qwen3-VL Family Behavior.

C. Evaluation Prompts

C.1. Prompt Templates

Tab. 4 presents the unified prompt template employed for all evaluations conducted on GroundingME. Tab. 5 details the prompt template used for the MLLM judge baseline during the best-of-N selection in our test-time scaling analysis. Tab. 6 shows the prompt template used for the text-only LLM judge to select the optimal response based on thinking trajectory during our test-time scaling analysis.

C.2. Prompt Robustness

We use 5 distinct prompts to evaluate Qwen3-VL-8B and Llama-Nemotron and yield accuracies of 31.36±2.02 and 10.07±0.44, matching previous scores and showing little variance. We also find 98.61% of Qwen3-VL-8B responses follow the instructed format, and fix the remaining by comprehensive rules to minimize instruction-following errors.

D. Human Rejection Verification

Given the poor performance of models on the Rejection category, we conduct a human verification study to validate the correctness and quality of our data. We randomly sample 100 instances from GroundingME for human binary classification (Reject/Non-Reject) annotation. To mitigate the risk of human annotators taking linguistic shortcuts based on distinct description styles, we restrict the sampling to the Discriminative and Rejection categories only, as their referring expressions exhibit structural similarity. The final sampled set includes 51 instances from the Rejection category. Considering Rejection as the positive class, human annotators achieve an Accuracy of 91%, with a Precision of 88.24%, a Recall of 93.75%, and an F1-score of 90.91%.

E. Commercial Model Notes

Regarding the evaluation of commercial models, we make a specific adjustment for the Gemini-2.5 series: we modify the required coordinate format in the prompt template (Tab. 4) to [y1, x1, y2, x2]. This modification is implemented because we observe that Gemini-2.5 is significantly more receptive to this output format, resulting in a measurable improvement in accuracy.

We do not report the evaluation results for GPT-5, Claude-Sonnet-4.5, and Grok-4 due to issues with their output. From cases in Tab. 7, we observe that the coordinates produced by these models using the unified prompt template (Tab. 4) suffered from substantial displacement and distortion, regardless of whether the output is interpreted as absolute pixel coordinates (red bounding box) or 0-999 normalized relative coordinates (blue bounding box). Furthermore, we fail to find an alternative coordinate format that yields usable results for these models.

F. Tool Use Results

We also conduct an evaluation of Claude-Sonnet-4.5 utilizing PyVision [1] for tool use. The total accuracy is 12.4%. The detailed subtask breakdown is as follows: Discriminative: 19.1% (App.: 15.4%, Cmp.: 32%, Txt.: 10%, Sta.: 19.2%); Spatial: 13.3% (Rel.: 13.3%, Cnt.: 13.3%); Limited: 9.0% (Occ.: 10.7%, Sml.: 7.3%); and Rejection: 9.5% (App.: 10%, Cmp.: 7.8%, Txt.: 14%, Sta.: 6%).

We assume that the model's ability to utilize tool use for multi-step cropping and magnification of 8K image to localize tiny objects should yield improved performance on the Limited_Small subcategory. However, the observed accuracy (7.3%) falls below our expectations. Through case study in Tab. 8, we find that subtle offset of bounding box size and position is a significant contributing factor to this unsatisfactory result.

G. Examples for Each L-2 Subcategory

In Tab. 9 through Tab. 20, we provide precise task definitions for all twelve L-2 subcategories in GroundingME, and present one representative example for each. In all displayed examples, the red bounding box indicates the correct ground-truth object.

References

[1] Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. Pyvision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025. 2

Table 3. Evaluation results across different IoU thresholds on GroundingME. All settings and abbreviations are the same as in §4.

Model	Discriminative			Spatial			Limited			Total		
	$\text{Acc}_{0.75}$	$\text{Acc}_{0.9}$	mAcc	$\text{Acc}_{0.75}$	$\text{Acc}_{0.9}$	mAcc	$\text{Acc}_{0.75}$	$\text{Acc}_{0.9}$	mAcc	$\text{Acc}_{0.75}$	$\text{Acc}_{0.9}$	mAcc
Phi-4-Multimodal	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.1
Llama-4-Maverick	8.8	0.5	9.5	8.3	0.7	9.8	1.7	0.0	1.9	6.0	1.5	6.6
Llama-4-Scout	7.4	0.5	8.0	3.7	0.0	4.9	1.3	0.0	1.4	3.5	0.6	4.0
LLaVA-O.V.-1.5-8B	2.9	0.5	4.0	0.7	0.0	1.1	0.0	0.0	0.7	0.8	0.1	1.4
Mistral-3.2-24B	0.0	0.0	1.1	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.5
Gemma-3-27B	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
Llama-Nemotron-8B	14.7	3.9	14.9	2.7	0.7	2.9	4.7	2.0	5.1	6.3	2.7	6.5
MiniCPM-V-4.5	1.0	0.0	2.4	0.7	0.0	1.4	0.0	0.0	1.1	0.4	0.0	1.2
InternVL3.5-8B	2.0	1.5	3.2	2.3	0.0	2.2	0.7	0.3	0.9	1.6	0.7	1.9
InternVL3.5-A28B	16.2	6.4	16.4	14.3	3.0	13.9	6.7	3.3	7.6	9.6	3.2	9.8
Keye-VL-1.5-8B	8.3	0.5	9.8	1.3	0.0	2.6	0.7	0.3	1.6	2.3	0.2	3.2
MiMo-VL-7B-RL	33.8	13.2	30.6	14.0	5.0	12.8	8.0	3.0	8.1	13.4	5.1	12.5
GLM-4.5V	44.1	32.4	42.5	37.7	24.7	34.5	19.0	8.3	18.2	26.0	16.5	24.5
Qwen2.5-VL-7B	24.0	12.3	22.4	10.3	4.0	9.3	9.3	1.3	8.4	10.8	4.2	9.9
Qwen2.5-VL-32B	38.7	15.2	33.6	29.0	13.0	27.2	9.7	3.0	10.8	19.4	7.9	18.2
Qwen2.5-VL-72B	42.2	21.1	37.4	30.0	14.7	28.4	14.3	4.3	14.4	22.4	10.5	21.0
Qwen3-VL-2B	39.2	31.9	37.8	10.3	8.0	9.9	15.3	7.0	16.5	15.6	10.9	15.6
Qwen3-VL-4B	52.9	43.1	50.2	25.7	20.0	24.3	29.7	11.7	29.4	27.3	18.2	26.2
Qwen3-VL-8B	57.4	47.1	55.0	23.7	21.0	23.3	27.3	16.7	26.5	26.9	20.8	26.0
Qwen3-VL-32B	71.1	54.9	65.9	44.0	32.7	41.3	28.3	19.0	26.3	36.0	26.6	33.6
Qwen3-VL-A3B	61.3	50.5	57.8	27.3	20.0	25.6	32.0	15.3	31.4	30.1	20.8	28.7
Qwen3-VL-A22B	66.7	54.9	63.3	46.3	38.3	44.1	37.3	21.7	36.7	38.5	29.1	37.0
Gemini-2.5-Pro	22.5	11.8	22.9	23.0	12.3	21.7	2.7	0.7	3.5	13.6	7.7	13.6
Gemini-2.5-Flash	27.5	14.7	26.6	19.0	13.3	19.0	6.3	1.0	7.2	13.1	7.3	13.2
Seed-1.6-V.-250815	55.4	41.7	51.7	51.3	35.7	48.0	30.7	19.0	30.1	35.9	25.0	34.0

<image>

All spatial relationships are defined from the viewer's perspective, where 'front' means closer to the viewer and 'back' means farther from the viewer. Please provide the bounding box coordinate of the object the following statement describes:

{description}

Ensure that all details mentioned about the object are accurate. Provide at most one bounding box. If a matching object is found, provide its bounding box as a JSON in the format {"bbox_2d": [x1, y1, x2, y2]}. If no matching object is found, output {"bbox_2d": null}.

Table 4. Prompt template for all evaluations on GroundingME.

<image>

Role and Task

You are an expert-level Visual Grounding Adjudicator. Your task is to evaluate two proposed bounding boxes (Bbox A and Bbox B) for a given image and user instruction, and determine which one is the more accurate and superior choice.

Input

Instruction: {instruction}
Bbox A: {bbox_a}
Bbox B: {bbox_b}

Output
Explain your reasoning, then conclude with your final choice in the format \boxed{A} or \boxed{B}.

Table 5. Prompt template for multimodal judge models in test-time scaling analysis.

Role and Task

You are a rigorous AI reasoning process analyst. Your task is to compare the two responses provided (Response A and Response B) based on the five principles below and select the superior one.

Core Evaluation Principles

All of your judgments MUST be strictly based on the following five points:

1. **Instruction Understanding**: Evaluate whether the model has correctly and comprehensively understood the description in the user's instruction, including all details, constraints, and limitations.
2. **Visual Observation**: Evaluate whether the model has comprehensively and meticulously observed the image, identifying as many objects and their attributes or spatial relationships as possible.
3. **Logical Reasoning**: Evaluate whether each step of the reasoning is logical and free of contradictions, fallacies, or unsubstantiated leaps.
4. **Analytical Rigor**: Evaluate whether the conclusion was reached hastily or formed after carefully analyzing and comparing multiple possibilities.
5. **Conclusion Support**: Evaluate whether the final answer is strongly supported and uniquely derived from the thought process, rather than being disconnected from it.

Input

[Original Task Instruction]
{instruction}

[Full Content of Response A]
{response_a}

[Full Content of Response B]
{response_b}

Output

Your final selection must be **only** one of the following two lines, with no other text before or after: \boxed{A} or \boxed{B}.

Table 6. Prompt template for text-only judge models in test-time scaling analysis.

Description:
This is a small, light green plastic stool with a top and four tapered legs that splay slightly outwards. The top surface has a subtle pattern. A small sticker is attached to it. Its size suggests it's a common, lightweight outdoor seating option.
Correct Answer: [544, 1102, 940, 1498]

GPT-5 Answer: [320, 600, 520, 760]

Claude-4.5 Answer: [89, 632, 301, 869]

Grok-4 Answer: [59, 322, 130, 410]

Table 7. Cases of outputs from unreported commercial models.

Description:
The object is a young girl. She is squatting on a wooden raft or platform by the water.

Original Image: width=7680, height=5046

Correct Answer: [3389, 3448, 3487, 3530]

Claude-4.5 Answer: [0.4323, 0.6877, 0.4544, 0.7134]

Table 8. Case of Claude-Sonnet-4.5 output with tool use.

Subcategory 1: Discriminative_Appearance

Definition: Distinguishing objects based on subtle visual attributes like color or texture.

Description:
This is a white cube, likely a Mahjong tile, with a smooth, reflective surface. **On its top face, there are two blurry, dark vertical markings, which appear to be thin lines or abstract shapes, rendered out of focus.** The material seems to be a hard, glossy substance like plastic or ceramic.

Table 9. An example of Discriminative_Appearance Subtask.

Subcategory 2: Discriminative_Component

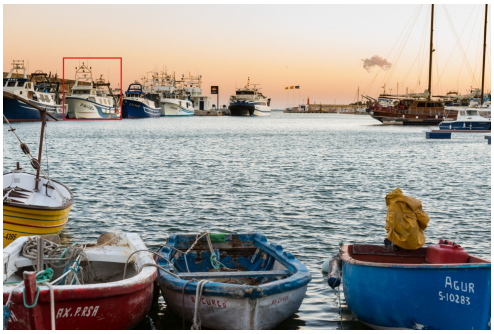
Definition: Distinguishing targets based on the presence or absence of a specific structural component.

Description:
This is a tall, slender tree with a relatively straight, thin trunk and sparse, upright branches. The trunk and branches are primarily dark brown to reddish-brown, suggesting sparse foliage. The texture appears rough and natural. **There are a few brown leaves concentrated on some of the upper branches**, indicating it might be a deciduous tree in a dry season or a tree with naturally sparse foliage.

Table 10. An example of Discriminative_Component Subtask.

Subcategory 3: Discriminative_Text

Definition: Distinguishing targets based on textual information embedded within the image.



Description:
This object is a boat. T**he number '1-1-03' is visible on its hull.**

Table 11. An example of Discriminative_Text Subtask.

Subcategory 4: Discriminative_State

Definition: Distinguishing objects based on their dynamic or static condition.

Description:
The jet is dark-colored, appearing black or very dark navy, with a sleek, aerodynamic design characteristic of a military training aircraft. **The jet is positioned vertically. It is actively emitting a vibrant, opaque red smoke trail from its rear.** The surface appears smooth and metallic.

Table 12. An example of Discriminative_State Subtask.

Subcategory 5: Spatial_Relationship

Definition: Grounding the target based on its spatial position relative to other entities.

Description:
The object is a hut. **Immediately to the left of this hut is another beach hut, which is light grey in color. To its immediate right is a vibrant blue beach hut.** Below the hut is a sturdy concrete wall or barrier, and further down is the sandy beach. Directly above and behind the hut is a lush green hillside covered with dense vegetation.

Table 13. An example of Spatial_Relationship Subtask.

Subcategory 6: Spatial_Counting

Definition: Grounding the target based on explicit quantitative or ordinal information within the scene.

Description:
The flag is made of fabric with creases as it moves in the breeze. **To its right, there are five more flags.**

Table 14. An example of Spatial_Counting Subtask.

Subcategory 7: Limited_Occlusion

Definition: Localizing objects with partial visibility caused by occlusion or truncation by the image frame.

Description:
The object is a traffic cone, **the first one from the right.**

Table 15. An example of Limited_Occlusion Subtask.

Subcategory 8: Limited_Small

Definition: Localizing objects with diminutive scale in ultra-high resolution images.

Description:
The object is a person holding a camera. It appears to be capturing a photograph while seated outside.

Table 16. An example of Limited_Small Subtask.

Subcategory 9: Rejection_Appearance

Definition: Rejecting the query due to a factual contradiction in the described visual attributes like color or texture.

Description:
The object is an elongated, oval-shaped foil balloon, **primarily red with a prominent vertical white stripe running down its center. On either side of the white stripe, there are yellow, circle shapes outlined with red patterns.** The balloon has a smooth, reflective texture typical of Mylar balloons.

Table 17. An example of Rejection_Appearance Subtask.

Subcategory 10: Rejection_Component

Definition: Rejecting the query due to a factual contradiction concerning a specific structural component.

Description:
The object is a young child, consisting of their bare legs and small feet. **The child wears a pair of pink flip-flop with a white sole on both feet.**

Table 18. An example of Rejection_Component Subtask.

Subcategory 11: Rejection_Text

Definition: Rejecting the query due to a factual mismatch with embedded textual information.

Description:
The object is a white VGA coaxial cable coiled inside a clear plastic blister pack with a blue backing. **A green price label with black text "180" is affixed to the front of the packaging.** The cable itself has a smooth appearance and is neatly coiled into a circular shape.

Table 19. An example of Rejection_Text Subtask.

Subcategory 12: Rejection_State

Definition: Rejecting the query because the object's described dynamic or static condition is factually incorrect.

Description:
The object is a maroon-colored compact SUV with its doors closed. Its front end is heavily damaged and crushed, indicating an impact. The paint on the undamaged parts of the car appears somewhat glossy, and the windshield and windows are visible.

Table 20. An example of Rejection_State Subtask.