

MM-Spatial: Exploring 3D Spatial Understanding in Multimodal LLMs

Supplementary Material

A. More Details about the CA-VQA Data

A.1. Spatial Task Categories

We here provide more details on the different spatial task categories covered in CA-VQA, with visualizations of examples provided in Figs. 5 to 7.

- **Binary.**
 - **Viewpoint-Dependent.** We consider the spatial relationships *left vs. right* and *in front vs. behind* between two objects, as determined from the current camera pose / viewpoint:²
 - * *Left vs. Right.* We determine the answer based on the horizontal coordinates of the objects' 2D bounding box centers.
 - * *In front vs. Behind.* We determine the answer based on the distances between the camera and the objects' 3D bounding box centers.
 - **Relative Object Size.** We determine the answer based on the objects' *width*, *length* or *height*, as defined in **Regression** below.
 - **Object Presence.** For each sample asking about an object present in the image, we also generate a negative sample which asks about a (randomly sampled) object *not* present in the image, to ensure a uniform distribution over answers (*Yes / No*).
- **Counting.** We determine the answer by simply counting the number of bounding boxes present in the image for a given object class. We also generate negative samples with (randomly sampled) objects not present in the image (i.e., such that the correct answer is 0).
- **Multi-choice.** This covers questions across the other spatial task categories, except for 2D and 3D grounding. We randomize the order of the options, obtaining the incorrect options as follows:
 - **Regression (Metric Estimation).** We compute three wrong options with either 10% increments deviating from the real answer, or 5cm, whichever value is larger.
 - **Counting.** We always ensure that 0 is an option (i.e., object is not present). We then randomly sample (additional) wrong options among the non-zero integers within $[GT - 3, GT + 3]$ (where GT is the correct answer), s.t. the total number of options is 4.
 - **2D / 3D Referring.** We randomly sample three wrong

²Note that we do not consider *above vs. below* to avoid ambiguity: "above" could either refer to 2D image space (i.e., the 2D bounding box of A is above that of B), or to 3D space, where the latter can be ambiguous as well (i.e., do we just require that the 3D bounding box of A is located higher in terms of vertical dimension, or do we also require that A is located directly above B in terms of horizontal dimensions – the latter might best match with how humans colloquially define "above").

object classes.

- **Regression (Metric Estimation).**
 - **Egocentric Distance.** The distance between the camera and the *closest* point of the object's point cloud.
 - **Object Distance.** We consider both (minimum) distance and center distance between two objects:
 - * *(Minimum) Distance.* The distance between the *closest* points of the two objects' point clouds (i.e., minimum point distance).
 - * *Center Distance.* The distance between the *center* points of the two objects' 3D bounding boxes.
 - **Object Size.** We consider the 3D dimensions *width*, *length* and *height*, defined as follows:
 - * *Width.* The length of the *larger* horizontal edge of the object's 3D bounding box (i.e., $\max(x_{len}, z_{len})$).
 - * *Length.* The length of the *shorter* horizontal edge of the object's 3D bounding box (i.e., $\min(x_{len}, z_{len})$).
 - * *Height.* The length of the vertical edge of the object's 3D bounding box (i.e., y_{len}).
- **2D Grounding.** We use the 2D bounding box obtained from projecting the object's 3D bounding box into 2D image space.
- **3D Grounding.** We directly use the 3D bounding boxes provided in CA-1M [61].

A.2. Depth: Chain-of-Thought (CoT) / Tool-Use

We prepare multi-step CoT responses involving depth for questions within the *Binary* (only "behind vs. in front") and *Regression (Metric Estimation)* categories, as the ground truth answers for those rely on depth information. We also did preliminary experiments with *3D Grounding* samples, but found that performance does not improve / even slightly regresses there, so we did not include any such samples in the final dataset.³

The sequence format of the samples is illustrated in Fig. 3, involving the target objects' 2D bounding boxes and depth values and the final (original) answer. We use the GT depth maps for generating the training examples, extracting the median depth value within the object's 2D bounding box⁴. At test time, we then consider two alternative approaches for obtaining the depth values:

³We hypothesize that 3D grounding is too complex of a task to benefit from the simple depth information provided in the multi-step CoT answers, and that the model might just get confused. We leave a more comprehensive study of how to benefit 3D grounding with CoT for future work.

⁴We also did preliminary experiments with other ways to extract a single depth value from the depth map within the 2D bounding box, such as the value at the center of the box or percentiles other than the median, but did not see significant improvements over using the median, which we found to be a robust choice.

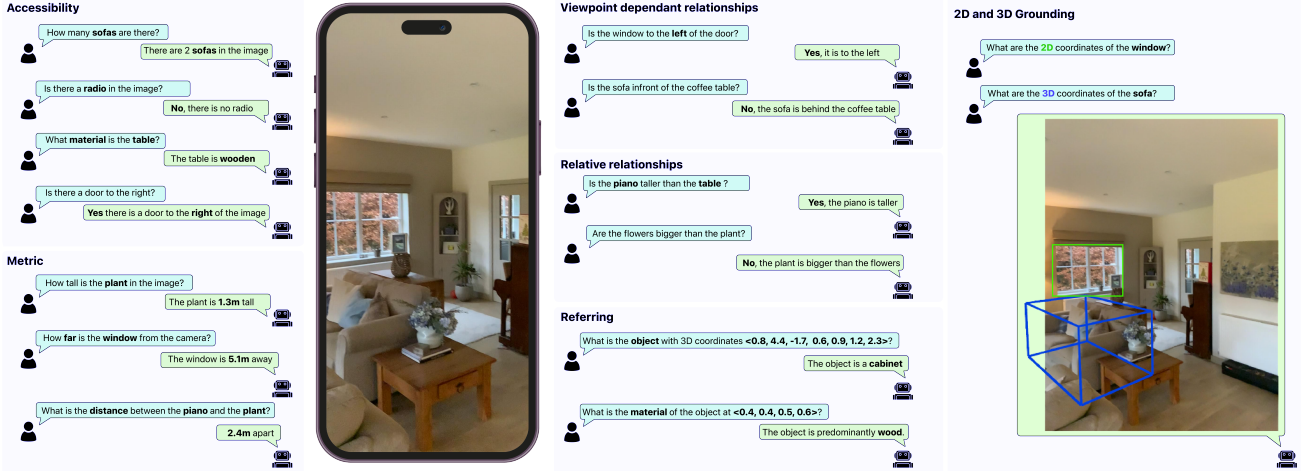


Figure 5. **CA-VQA Overview.** Example QA pairs from our Cubify Anything VQA (CA-VQA) dataset, aiming to unlock object-centric 3D spatial understanding in MLLMs. Using high-quality 3D ground truth annotations from CA-1M [61], we generate spatial perception questions across a variety of different tasks, e.g., involving **relative relationships**, **metric measurements**, and **3D object bounding boxes**.

- **Model prediction (CoT).** We let the model predict the depth values (called *CoT* in the experiments). As the model was trained on sequences involving the ground truth depth values, the models learn to predict depth. Our experiments reveal the accuracy of the resulting depth estimates.
- **Tool-use.** We allow the model to leverage a given depth map via tool-use. I.e., for a function call of the form $\text{Depth}(\text{bbox}) \rightarrow$ we extract the median depth value within the 2D bounding box, insert the depth value into the sequence, and then let the model continue its prediction to arrive at the final answer (see Fig. 3).

B. Optimal Data Mixture for MM-Spatial

We aim to build a generalist MLLM that excels across a variety of diverse tasks – as opposed to a specialist that *only* excels at spatial understanding. To this end, we identify the mixture weight for the new spatial data that achieves the best performance trade-off between the spatial vs. all other benchmark categories: general, knowledge, text-rich, 2D referring & grounding. Investigating the effect of adding a new model capability is particularly relevant for models with limited capacity, such as the 3B model we consider.

Results are shown in Tab. 7. MM-Spatial maintains similar performance as the MM1.5 baseline across most task categories, while significantly improving on the Spatial category. This suggests that our model can successfully adopt the new spatial understanding capability without regressing on all the other capabilities, resulting in a generalist MLLM. The data mixture ratio of 2:1 (spatial:general) provides a good performance trade-off and is used for MM-Spatial throughout. We also consider a spatial *Specialist Model* that

is trained on CA-VQA only; however, this model provides only a small improvement on the spatial category, while regressing substantially on all other benchmark categories. We use specialist models for some of our ablations to speed up experimentation. Appendix C shows the detailed result breakdowns across the different task categories, compared to SOTA models.

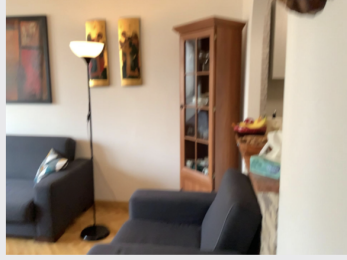

C. Results on Further Benchmark Categories

We here present a more detailed analysis of MM-Spatial compared with SOTA baselines across the different benchmark categories. Results on general and knowledge benchmarks are shown in Tab. 8, results on text-rich benchmarks are shown in Tab. 9, and results on 2D referring & grounding benchmarks are shown in Tab. 10. Overall, we observe that our MM-Spatial model maintains a level of performance similar to the vanilla MM1.5 baseline. This suggests that our model is able to successfully adopt the new spatial understanding capability without sacrificing performance on all the other model capabilities, resulting in a generalist MLLM.

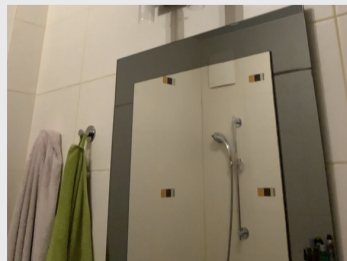

D. Analysis of Blind Filtering Procedure

Tab. 11 analyses the effectiveness of our blind filtering procedure outlined in Sec. 3.1 in ensuring that our CA-VQA benchmark becomes more reliant on vision input. This is in contrast to some of the tasks from the other spatial understanding benchmarks we consider (CV-Bench and SpatialRGPT-Bench), where we found that blind models can perform very strongly and even rival models with vision input in some cases (see Sec. 5). Hence, these benchmarks would likely also benefit from blind filtering.

Binary

	<p>Viewpoint-Dependent</p> <p>Q: Is the cabinet in front of the tissue box?</p> <p>A: No</p> <p>Relative Object Size</p> <p>Q: Is the cushion higher than the tissue box?</p> <p>A: Yes</p> <p>Object Presence</p> <p>Q: Is there a watering can in the image?</p> <p>A: No</p>		<p>Viewpoint-Dependent</p> <p>Q: Is the glass candle to the right of the rug?</p> <p>A: No</p> <p>Relative Object Size</p> <p>Q: Is the sofa shorter than the radiator?</p> <p>A: No</p> <p>Object Presence</p> <p>Q: Is there a cushion in the image?</p> <p>A: Yes</p>
---	--	--	---

Counting

	<p>Q: How many mirrors are in the image?</p> <p>A: 1</p> <p>Q: How many towels are in the image?</p> <p>A: 2</p> <p>Q: How many goggles are in the image?</p> <p>A: 0</p>		<p>Q: How many bottles are in the image?</p> <p>A: 2</p> <p>Q: How many toilet seats are in the image?</p> <p>A: 1</p> <p>Q: How many drill machines are in the image?</p> <p>A: 0</p>
---	---	--	--

Multi-choice




	<p>Metric Estimation</p> <p>Q: How high is the basket?</p> <p>A. 13cm B. 8cm C. 18cm D. 23cm</p> <p>A: C</p> <p>Counting</p> <p>Q: How many glasses are in the image?</p> <p>A. 2 B. 0 C. 3 D. 1</p> <p>A: A</p>		<p>2D Referring</p> <p>Q: What object has the 2D image coordinates $<0.51, 0.30, 0.92, 0.67>$?</p> <p>A. cabinet B. chair C. plaque D. heater</p> <p>A: D</p> <p>Relative Relationship</p> <p>Q: Is the oven wider than the towel?</p> <p>A. no B. yes</p> <p>A: B</p>
---	--	--	---

Figure 6. Examples of CA-VQA data samples from the Binary, Counting and Multi-choice categories.

E. Axis-aligned vs. Oriented 3D Boxes

Fig. 8 emphasizes the fundamental difference between axis-aligned (AABB) and oriented (OBB) 3D bounding boxes and how they affect the resulting object dimensions. This provides an indication of the misalignment issues arising when evaluating a model trained on data based on OBB ground truth (i.e., MM-Spatial, which is based on the gravity-aligned 7-DOF yaw-oriented 3D bounding boxes from CA-1M) on a benchmark based on AABB ground truth (i.e., SpatialRGPT-Bench) and vice versa (i.e., evaluating SpatialRGPT on CA-VQA), as seen in Secs. 5.3 and 5.5.


Regression (Metric Estimation)



Egocentric Distance
Q: How far away is the wall clock?
A: 1.49m

Object Distance
Q: What is the distance between the fruit bowl and the table?
A: 58cm

Object Size
Q: How high is the fruit bowl?
A: 7cm




Egocentric Distance
Q: How far away is the cabinet?
A: 2.03m

Object Distance
Q: What is the distance between the center of the cabinet and the center of the basket?
A: 1.17m


Object Size
Q: How long is the basket?
A: 37cm

2D Grounding



Q: What are the 2D image coordinates of the kitchen island?
A: <0.138,0.347,0.508,0.525>

Q: What are the 2D image coordinates of the christmas tree?
A: <0.558,0.209,0.803,0.472>



Q: What are the 2D image coordinates of the object with primary material paper?
A: <0.621,0.422,0.692,0.475>

Q: What are the 2D image coordinates of the stove?
A: <0.0,0.497,0.225,0.688>

Figure 7. Examples of CA-VQA data samples from the Regression (Metric Estimation) and 2D Grounding categories.

Model	Benchmark Category Averages										
	Mix. Ratio		Spatial Understanding				General	Knowledge	Text-rich	Refer&Ground	Avg.
	Rel.	Eff.	CA-VQA	CV-Bench	SRGPT-Bench	Avg.					
MM1.5-3B [128]	0:1	0:100	28.9	64.9	26.0	39.9	64.7	46.2	62.1	77.7	58.1
MM-Spatial-3B	1:1	12:88	66.3	91.2	52.8	70.1	65.0	46.2	62.1	79.1	64.5
	2:1	22:78	67.1	92.4	53.7	71.1	64.8	46.7	61.4	78.8	64.5
	4:1	36:64	67.3	93.0	52.7	71.0	65.0	44.9	60.7	78.0	63.9
	8:1	54:46	67.4	93.1	53.7	71.4	64.8	46.8	61.2	79.0	64.6
MM-Spatial-3B	1:0	100:0	67.1	93.0	54.1	71.4	42.6	34.7	17.2	23.9	38.0

Table 7. **Data Mixture Ratio Results.** Comparison of different data mixture ratios – both (Rel)ative to the *General* category (as in MM1.5), and (Eff)ective when considering the dataset sizes – on aggregated metrics across the different benchmark categories. Overall, MM-Spatial is a generalist MMLM that improves a lot on the *Spatial* category while maintaining strong performance on the other categories. The data mixture ratio of 2:1 (spatial:general) provides a good performance trade-off and is used for MM-Spatial throughout. The last line considers a spatial *Specialist Model* that is trained on CA-VQA only; this model provides only a minor improvement on the spatial category, while regressing substantially on all other benchmark categories.

Model	Knowledge Benchmarks			General Benchmarks					
	AI2D (test)	MMMU (val)	MathV (testmini)	MME (P/C)	SEED ^I	POPE	LLaVA ^W	MM-Vet	RealWorldQA
MiniCPM-V 2.0-3B [119]	62.9	38.2	38.7	1808.2 [†]	67.1	87.8	69.2	38.2	55.8
VILA1.5-3B [76]	–	33.3	–	1442.4/–	67.9	85.9	–	–	–
SpatialRGPT-VILA-1.5-3B [27]	–	33.0	–	1424.0/–	69.0	85.5	–	38.2	–
TinyLLaVA [137]	–	–	–	1464.9/–	–	86.4	75.8	32.0	–
Gemini Nano-2 [103]	51.0	32.6	30.6	–	–	–	–	–	–
Bunny [44]	–	41.4	–	1581.5/361.1	72.5	87.2	–	–	–
BLIP-3 [115]	–	41.1	39.6	–	72.2	87.0	–	–	60.5
Phi-3-Vision-4B [1]	76.7	40.4	44.5	1441.6/320.0	71.8	85.8	71.6	46.2	59.4
MM1.5-3B [128]	64.5	37.1	37.1	1423.7/277.9	70.2	87.9	74.3	37.1	57.7
MM-Spatial-3B	63.6	36.6	38.4	1530.5/251.8	71.3	88.0	69.9	38.0	59.0
Gemini-1.5-Pro [93]	79.1	60.6	57.7	2110.6 [†]	–	88.2	95.3	64.0	64.1
GPT-4V [86]	75.9	53.8	48.7	1771.5 [†]	71.6	75.4	93.1	56.8	56.5
GPT-4o [50]	84.6	69.2	61.3	2310.3 [†]	77.1	85.6	102.0	69.1	75.4

Table 8. **Knowledge and General Benchmark Results.** Comparison with SOTA models on knowledge and general benchmarks. (†) Sum of P and C scores. Gemini-1.5-Pro, GPT-4V and GPT-4o numbers are from [33].

Model	WTQ (test)	TabFact (test)	OCRBench (test)	ChartQA (test)	TextVQA (val)	DocVQA (val)	InfoVQA (val)
MiniCPM-V 2.0-3B [119]	24.2	58.2	60.5	59.8	74.1	71.9	37.6
TinyLLaVA [137]	–	–	–	–	59.1	–	–
Gemini Nano-2 [103]	–	–	–	51.9	65.9	74.3	54.5
BLIP-3-4B [115]	–	–	–	–	71.0	–	–
Phi-3-Vision-4B [1]	47.4	67.8	63.7	81.4	70.1	83.3	49.0
MM1.5-3B [128]	37.3	70.5	63.0	73.6	74.4	82.0	45.5
MM-Spatial-3B	36.2	71.0	60.0	75.0	75.3	82.7	43.7
Gemini-1.5-Pro [93]	–	–	75.4	87.2	78.7	93.1	81.0
GPT-4V [86]	–	–	64.5	78.5 [†]	–	88.4 [†]	–
GPT-4o [50]	–	–	73.6	85.7 [†]	–	92.8 [†]	–

Table 9. **Text-rich Benchmark Results.** Comparison with SOTA models on text-rich benchmarks. (†) Numbers are obtained from [63].

Model	RefCOCO (testA/B)	RefCOCO+ (testA/B)	RefCOCOg (test)	Flickr30k (test)	LVIS-Ref (box/point)
MiniCPM-v2-3B [119]	–	–	–	–	48.2/47.7
Phi-3-Vision-4B [1]	46.3 / 36.1	42.0 / 28.8	37.6	27.12	53.8/54.5
InternVL2 [26]	88.2 / 75.9	82.8 / 63.3	78.3	51.6	51.0 / 51.1
MM1.5-3B [128]	91.7 / 85.7	87.67 / 75.23	85.9	85.1	74.0 / 58.2
MM-Spatial-3B	92.2 / 85.9	88.3 / 76.8	86.8	85.1	75.9 / 58.5

Table 10. **2D Referring & Grounding Benchmark Results.** Comparison with SOTA models on 2D referring and grounding benchmarks.

Model	Eval Inputs	Regression (Metric Estimation)						Average	
		Binary	Count.	Multi-c.	Ego-Dist.	Obj.-Dist.	Obj.-Size		
		Acc	Acc	Acc	Acc @ 10% Relative Error (ℓ_1)				
<i>Before Blind Filtering</i>									
①	GPT-4 [2]	Text	57.9	35.1	52.7	8.9	8.2	17.0	30.0
②	GPT-4V [86]	Image + Text	61.6	68.1	63.2	6.4	8.4	19.7	37.9
③	Improvement from using vision = ② − ①		+3.7	+33.0	+10.5	-2.5	+0.2	+2.7	+7.9
④	MM-Spatial-3B (Specialist)	Text	69.3	69.5	77.6	12.9	11.0	25.2	44.3
⑤	MM-Spatial-3B (Specialist)	Image + Text	83.8	76.9	84.2	46.9	25.4	29.5	57.8
⑥	Improvement from using vision = ⑤ - ④		+14.5	+7.4	+6.6	+34.0	+14.4	+4.3	+13.5
<i>After Blind Filtering</i>									
⑦	GPT-4 [2]	Text	9.6	8.5	9.6	6.2	6.2	5.8	7.7
⑧	GPT-4V [86]	Image + Text	39.2	63.3	32.9	11.4	9.3	10.1	27.7
⑨	Improvement from using vision = ⑧ − ⑦		+29.6	+54.8	+23.3	+5.2	+3.1	+4.3	+20.0
⑩	MM-Spatial-3B (Specialist)	Text	34.3	60.8	60.7	10.1	8.4	17.9	32.0
⑪	MM-Spatial-3B (Specialist)	Image + Text	69.6	73.3	77.4	47.3	24.4	24.3	52.7
⑫	Improvement from using vision = ⑪ − ⑩		+35.3	+12.5	+16.7	+37.2	+16.0	+6.4	+20.7
<i>Increase in Vision Improvement: Before vs. After Blind Filtering</i>									
⑬	GPT-4/V	= ⑨ − ③	+25.9	+21.8	+12.8	+7.7	+2.9	+1.6	+12.1
⑭	MM-Spatial-3B (Specialist)	= ⑫ − ⑥	+20.8	+5.1	+10.1	+3.2	+1.6	+2.1	+7.2

Table 11. **CA-VQA Blind Filtering Analysis.** We study how the improvement from using vision (i.e., comparing a vision-evaluated model vs. a blind-evaluated model) changes after applying the blind filtering strategy outlined in Sec. 3.1, which follows [25]. Our results confirm that after applying our filtering strategy, 1) blind models perform substantially worse, and 2) vision improvements (i.e., the delta between vision and blind models) increase substantially, for both GPT-4/V and MM-Spatial. This highlights the effectiveness of our blind filtering procedure in ensuring that our CA-VQA benchmark becomes more reliant on vision input (i.e., less susceptible to a strong language prior).



Figure 8. Comparative visualization of axis-aligned vs. oriented 3D bounding boxes, taken from the SpatialRGPT paper [27, Appendix K, Figure 11]. The object dimensions computed from AABBs can differ substantially from those computed from OBBs, depending on the object’s rotation. For sake of illustration, assume that the sofa is 2m wide and 0.8m deep. We then obtain the following altered object dimensions when using an AABB instead of an OBB, at different yaw rotation angles (i.e., considering 7-DOF bounding boxes that are gravity-aligned / parallel to the ground, as in CA-1M / CA-VQA): width ≈ 2.1 m and depth ≈ 1.7 m with 30° rotation; width ≈ 1.7 m and depth ≈ 2.1 m with 60° rotation; and width = 0.8m and depth = 2m with 90° rotation (i.e., “full” rotation resulting in swapped dimensions).