Think with 3D: Geometric Imagination Grounded Spatial Reasoning from Limited Views

Supplementary Material

In this supplementary material, we provide more technical details and experimental results, including 1) Detailed descriptions of training dataset and benchmarks in Sec. 1; 2) Additional results including benchmarking Qwen3-VL and LLaVA-OneVision-1.5, additional baselines, general image understanding, prompt-relevant region, as well as ablation studies on projection methods and 3D loss in Sec. 2; 3) Training curve for two stages of *3DThinker* in Sec. 3; 4) Details of the dataset generation process in Sec. 4; 5) Explanation of the claim in Sec. 5; 6) Cost in Sec. 6; and 7) More "think with 3D" visualization including representative and failure cases in Sec. 7.

1. Dataset

1.1. Training Dataset

MindCube-Only Training: Following [12], we utilize a subset of 10,000 question-images-answer pairs for the training phases in Tab.1 of the main text. That is, in stage 1, we use the 10,000 samples to construct the CoT data for supervised training. In stage 2, we utilize only the question–image–answer pairs of these 10,000 samples for reinforced training. *Notably, no annotated cognitive maps are used in any stage of our training.*

Large-Scale Curated Dataset Training: We further select data from MM-Spatial [2], SPAR [13], MindCube [12], and SpatialLadder [8] to construct a larger-scale dataset. We utilized Gemini 2.5 Pro to evaluate the data quality and balanced the distribution between single-image and multi-image categories, ultimately retaining 237K samples. The methodology for constructing CoT data from standard QA pairs is detailed in Sec. 3.1 in the main text and Sec. 4. Similarly, we employed the complete CoT dataset in stage 1, whereas stage 2 solely utilized the question-image-answer pairs. For the training process, stage 1 was trained on the entire dataset for one epoch, while stage 2 was trained on a sampled subset of 57K instances.

1.2. Benchmarks

MindCube-Tiny [12] contains 1,050 data, serves as the benchmark for evaluation of spatial understanding from limited views. The dataset includes three types of camera movements: Rotation (the camera remains stationary while rotating to observe the surroundings), Around (the camera moves in a circular path around the evaluated objects), and Among (the camera moves in a circular path among the evaluated objects).

Ego3D-Bench [3] is a comprehensive benchmark consisting of over 8,600 question–answer pairs, specifically designed to evaluate the spatial reasoning capabilities of model in ego-centric, multi-view outdoor environments. The dataset encompasses five tasks—Absolute Distance Measurement (Dist.), Relative Distance Measurement (Rel.), Localization (Loc.), Motion Reasoning (Mot.), and Travel Time (Time), each formulated in both ego-centric (Ego) and object-centric (Obj.) settings.

VSI-Bench [10] consists of over 5,000 question–answer pairs derived from 288 real-world videos. These videos encompass a wide variety of environments—including residential areas, professional settings (e.g., offices and laboratories), and industrial sites (e.g., factories), and span multiple geographic regions.

SPBench [8] comprises two different tasks: SPBench-SI and SPBench-MV, corresponding to the single-image and multi-view types, respectively. SPBench-SI includes four categories: absolute distance, object size, relative distance, and relative direction, containing a total of 1,009 samples. SPBench-MV extends the evaluation by incorporating object counting tasks, comprising 319 samples.

CV-Bench [9] contains 2,638 manually inspected examples, each formulated as a natural-language question to assess the fundamental 2D and 3D understanding capabilities of VLMs. The 2D understanding tasks focus on spatial relationships and object counting, while the 3D understanding tasks involve depth ordering and relative distance.

SPAR-Bench [13] covers 20 diverse spatial tasks—including depth estimation, distance measurement, spatial relations, etc. The benchmark spans single-view, multi-view, and video settings, comprising a total of 7,207 manually verified question–answer pairs.

ViewSpatial-Bench [7] focuses on reasoning from alternative spatial frames of reference, requiring models to adopt another entity's viewpoint. It consists of over 5,700 question–answer pairs derived from more than 1,000 3D scenes. This benchmark evaluates the spatial localization capabilities, specifically assessing both egocentric (camera-centered) and allocentric (human-centered) viewpoints across five distinct task types.

MMSI-Bench [11] comprises 1,000 multiple-choice questions from over 120,000 images. The benchmark focuses on positional relationships (six pairwise combinations of camera, object, and region), attributes (measurement and appearance), and motion (camera and object).

2. Additional Results

2.1. Benchmarking Qwen3-VL and LLaVA-OneVision-1.5

We further conducted experiments on the Qwen3-VL and LLaVA-OneVision-1.5. Note that Qwen3-VL adopts <think> as a predefined special token, thus we have modified to <thinking> to avoid potential conflicts. On the other hand, *we modify the stage 2 from token-level GRPO [4] to sequence-level GSPO [14] for all Mixture-of-Experts(MoE)-based [6] models* to maintain stable training.

Qwen3-VL: Tab. 1 shows that our method consistently boosts the spatial understanding capability of Qwen3-VL. On the MindCube-Tiny [12] dataset, *3DThinker* achieves a maximum improvement of **191.5%** (75.8 vs. 25.0), elevating the VLM from random guessing to highly accurate understanding on spatial reasoning tasks. On the Ego3D-Bench [3] dataset, *3DThinker* yields up to a **29.9%** improvement (56.1 vs. 43.2), demonstrating strong generalization ability. Moreover, we validated our method on Qwen3-VL-30B-A3B sparse model as well, confirming the effectiveness of our method across both dense and sparse architectures.

LLaVA-OneVision-1.5: As shown in Tab. 2, we further extended our training to another foundation model. Specifically, on LLaVA-OneVision-1.5, *3DThinker* consistently yielded significant performance gains. Notably, it achieved a **68.07%** (63.7 vs. 37.9) improvement on MindCube-Tiny and a **7.5%** (48.7 vs. 45.3) increase on Ego3D-Bench.

2.2. Additional Baselines

As shown in Tab. 3, *3DThinker consistently outperforms previous methods*, surpassing the SOTA by **8.79%** (65.6 vs. 60.3). This further highlights the effectiveness of incorporating 3D mentaling for spatial reasoning.

2.3. General Image Understanding

As shown in Tab. 4, our method largely maintains, and on some benchmarks even substantially improves. For example, on POPE *3DThinker* outperforms the base model by **2.91%** (88.4 vs. 85.9). On MME, which predominantly consists of 2D image understanding tasks (e.g., OCR, numerical reasoning), our results are still comparable to the base model. These results suggest that *3DThinker* effectively transfers the base model's 2D image understanding abilities and *exhibits strong robustness*.

2.4. Prompt-Relevant Region

As shown in Fig. 1, we provide a quantitative result to substantiate the claim of "prompt-relevant". Specifically, we compute the point cloud density within the red 3D bounding box and compare it with the overall point cloud density. The resulting densities are **196494.67** vs. 1332.54 points/unit^3.

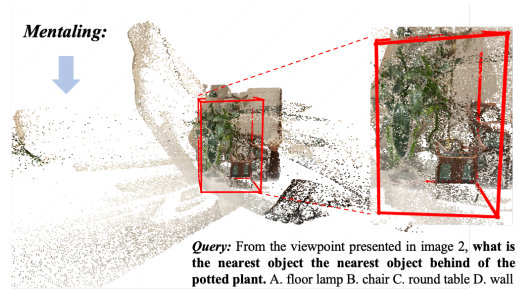


Figure 1. Density calculation for the Mentaled point cloud.

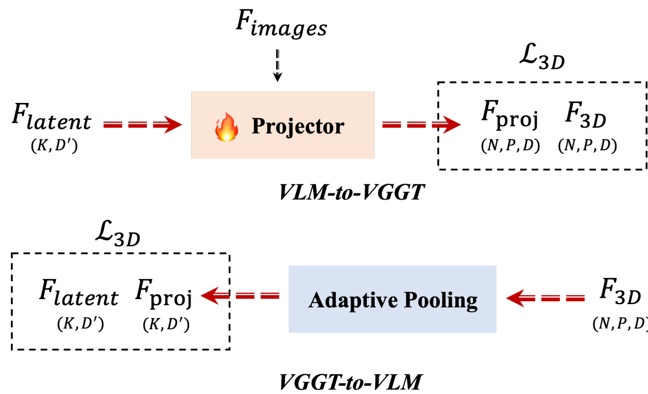


Figure 2. Illustration of two methods for projection. Above: project the latents generated by VLM into the feature space of the 3D foundation model; Below: project the features of the 3D foundation model into the feature space of the VLM.

2.5. Ablation on Projection Methods

As illustrated in Fig. 2, we present a more detailed comparison of the two feature alignment strategies. The above method requires training the projector, whereas the below one can be implemented directly via pooling [5] without any training. As discussed in Sec. 4.5, the two methods achieve comparable performance (74.1 vs. **75.2**). However, the above method not only maintains slightly better performance, but also *enables the recovery of 3D representations from the VLM latent space*. Therefore, we adopt the VLM-to-VGGT projection in our framework.

2.6. Ablation on 3D Loss

To evaluate the impact of $\mathcal{L}_{3D}$ on our framework, we further conduct the ablation where the 3D foundation model was removed, and only the cross-entropy loss ($\mathcal{L}_{\text{text}}$) was applied to tokens other than the 3D special token. As shown in Tab. 5, using Qwen2.5-VL-3B as an example, the performance drops from 62.7 (full) to 54.1 without $\mathcal{L}_{3D}$, though it remains slightly higher than the CoT SFT (53.4). These results highlight two key insights: (1) *incorporating the 3D foundation model helps the VLM utilize the special token to*

Table 1. Accuracy comparison between **LATEST** Qwen3-VL (under varying parameter scales and architectures) and *3DThinker* on different benchmarks, with our training conducted on stage 1 (S1) and on both stage 1 and stage 2 (S1 + S2). The best results achieved based on different VLMs are **bolded**. The overall/average results of each model are highlighted in blue.

Method	**MindCube-Tiny [12]**				**Ego3D-Bench [3]**								
	Rotation	**Among**	**Around**	**Overall ↑**	**Ego Dist.**	**Obj. Dist.**	**Loc.**	**Ego Mot.**	**Obj. Mot.**	**Travel Time**	**Ego Rel.**	**Obj. Rel.**	**Avg.↑**
Qwen3-VL-2B	29.0	32.2	46.8	35.0	24.7	23.9	30.4	58.7	56.4	32.8	57.4	61.1	43.2
3DThinker-S1$_{\text{Qwen3-2B}}$	40.5	67.5	77.2	64.7	45.9	41.6	36.0	60.6	59.2	37.2	62.0	72.1	51.8
3DThinker-S1+S2$_{\text{Qwen3-2B}}$	**53.0**	**83.2**	**78.4**	**76.2**	**53.1**	**51.3**	**36.4**	**61.4**	**62.8**	**38.2**	**65.3**	**80.5**	**56.1**
Qwen3-VL-4B	34.0	18.7	37.2	26.0	41.3	40.9	27.9	56.9	56.6	37.8	60.1	68.2	48.7
3DThinker-S1$_{\text{Qwen3-4B}}$	43.0	65.3	76.0	63.6	54.0	49.9	35.1	59.3	60.8	**40.8**	64.9	76.1	55.1
3DThinker-S1+S2$_{\text{Qwen3-4B}}$	**55.0**	**81.7**	**78.4**	**75.8**	**62.3**	**56.5**	**35.5**	**61.4**	**64.1**	40.4	**68.3**	**84.0**	**59.1**
Qwen3-VL-8B	31.5	31.0	41.2	33.5	38.9	32.6	31.6	45.3	56.8	40.8	62.4	69.3	47.2
3DThinker-S1$_{\text{Qwen3-8B}}$	42.0	67.8	76.8	65.0	52.1	45.9	37.1	51.5	61.8	**43.2**	67.8	78.2	54.7
3DThinker-S1+S2$_{\text{Qwen3-8B}}$	**56.5**	**82.5**	**78.8**	**76.7**	**63.0**	**56.8**	**38.1**	**55.5**	**66.2**	**43.2**	**71.1**	**86.1**	**60.0**
Qwen3-VL-32B	29.5	33.5	38.0	33.8	51.5	40.7	40.9	76.4	58.2	40.4	61.5	81.4	56.4
3DThinker-S1$_{\text{Qwen3-32B}}$	41.0	69.5	77.6	66.1	60.4	51.0	48.3	78.6	67.9	**43.2**	68.1	87.3	63.1
3DThinker-S1+S2$_{\text{Qwen3-32B}}$	**55.0**	**84.2**	**79.6**	**77.5**	**69.3**	**62.2**	**48.7**	**85.1**	**73.5**	43.0	**73.1**	**90.3**	**68.2**
Qwen3-VL-30B-A3B	37.0	43.9	45.6	42.9	49.1	45.8	28.4	64.5	60.0	34.7	63.3	73.3	52.4
3DThinker-S1$_{\text{Qwen3-30B-A3B}}$	46.5	72.8	82.0	70.0	60.1	54.4	38.3	72.2	70.3	**38.0**	71.0	80.1	60.6
3DThinker-S1+S2$_{\text{Qwen3-30B-A3B}}$	**60.0**	**86.7**	**83.2**	**80.8**	**69.4**	**65.1**	**39.1**	**79.3**	**75.1**	37.6	**76.3**	**88.0**	**66.2**

Table 2. Accuracy comparison of generalist VLMs and our method (*3DThinker*) on MindCube-Tiny and Ego3D-Bench, with our training conducted on stage 1 (S1) and on both stage 1 and stage 2 (S1 + S2). The best results achieved based on different VLMs are **bolded**. The overall/average results of each model are highlighted in blue, with the best results among all models highlighted in **red**.

Method	**MindCube-Tiny**				**Ego3D-Bench**								
	Rotation	**Among**	**Around**	**Overall ↑**	**Ego Dist.**	**Obj. Dist.**	**Loc.**	**Ego Mot.**	**Obj. Mot.**	**Travel Time**	**Ego Rel.**	**Obj. Rel.**	**Avg.↑**
LLaVA-OneVision-1.5 Family [1]													
LLaVA-OneVision-1.5-4B	33.5	38.0	49.2	39.8	39.7	37.1	29.2	51.4	51.8	**34.1**	52.4	73.5	46.2
3DThinker-S1$_{\text{LLaVA-O-1.5-4B}}$	41.5	59.8	**66.0**	57.8	40.0	39.1	33.1	51.1	**52.6**	30.9	58.6	**73.8**	47.4
3DThinker-S1+S2$_{\text{LLaVA-O-1.5-4B}}$	**48.0**	**67.5**	65.2	**63.2**	**40.2**	**39.9**	**34.2**	**51.9**	52.3	30.8	**61.8**	**73.8**	**48.1**
LLaVA-OneVision-1.5-8B	34.5	34.7	48.4	37.9	30.3	36.6	34.3	44.9	51.9	**36.9**	53.4	74.4	45.3
3DThinker-S1$_{\text{LLaVA-O-1.5-8B}}$	43.0	57.8	**64.8**	56.7	35.1	39.0	36.1	44.9	53.2	31.9	61.0	73.8	46.9
3DThinker-S1+S2$_{\text{LLaVA-O-1.5-8B}}$	**49.0**	**68.2**	**64.8**	**63.7**	**36.5**	**41.5**	**37.0**	**46.2**	**53.3**	32.8	**64.9**	**77.2**	**48.7**

Table 3. Performance comparison of all baselines on 3DSRBench.

SpatialReasoner	SpatialReasoner-R1	SpatialThinker	*3DThinker* *
60.3	55.7	56.4	**65.6**

Table 4. Results on general VLM benchmarks.

Method	**MME^{P}**	**MME^{C}**	**POPE**	**SEED-I**
Qwen2.5-VL-7B	1670	**623**	85.9	77.0
3DThinker *	**1677**	610	**88.4**	**78.9**

Table 5. Ablation study on 3D alignment loss on MindCube-Tiny in terms of Qwen2.5-VL-3B.

Method	**MindCube-Tiny**			
	Rotation	**Among**	**Around**	**Overall ↑**
w/o $\mathcal{L}_{3D}$	35.0	55.1	66.8	54.1
Full	**44.0**	**64.8**	**72.4**	**62.7**

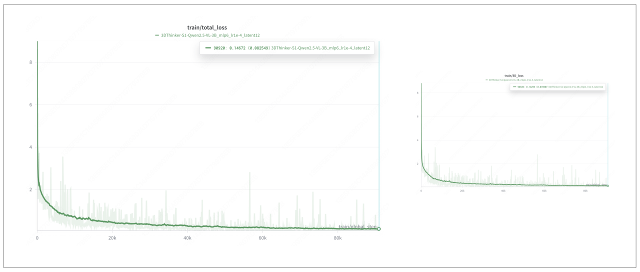
encode geometric information, thereby enhancing the spatial reasoning capability; and (2) even without the 3D foundation model, *introducing special tokens increases the representational flexibility of the model*, leading to moderate improvements on certain metrics.

3. Training Curve

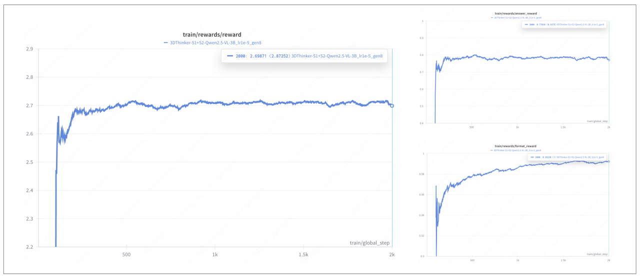
As shown in Fig. 3, we visualize the training curves of our two-stage process in terms of training on MindCube-only. The loss in stage 1 (green) converges after approximately 20k steps, reaching around 0.15. The reward in stage 2 (blue) converges much faster, stabilizing after roughly 500 steps with a reward value of about 2.7. In the bottom-right panel, we further plot the curves of the format reward (r_{format}) and the answer reward (r_{ans}). Both exhibit a brief decline during the early phase of training, followed by a steady upward trend. *This initial drop can be attributed to the model's exploration behavior at the beginning of reinforcement learning*, during which it searches the solution space and may temporarily follow suboptimal trajectories before discovering more optimal ones that lead to rapid performance improvement.

4. Details for Dataset Generation

For the training dataset, which consists of image set from different views I, question Q and GT response R, we employed GPT-4.1 (M) to complete the CoT



Stage 1: loss



Stage 2: reward

Figure 3. Visualization of our training curves in terms of Qwen2.5-VL-3B on stage 1 (green curves) and stage 2 (blue curves). Top-left: total loss of stage 1; Top-right: 3D loss of stage 1; Bottom-left: total reward of stage 2; Bottom-right: answer reward and format reward of stage 2.

> The Question, Images, and Answer are presented as follows. Please help me start with the mental 3D scene special token <output_3D> and complete the reasoning chain.
> Question: {question},
> Answer: {answer}.

Figure 4. Prompt for the CoT generation, with the input of question and GT answer.

> You are an advanced AI reasoning assistant with expertise in 3D spatial understanding. Your task is to complete the intermediate reasoning process using the provided Question, Images, and Answer. To assist your reasoning, you should mentally construct a 3D layout of the scene.
>
> Wrap your reasoning process with <think></think> tags and your final result with <answer></answer> tags. The mental 3D scene you imagine will be represented by the special token <output_3D>, which should be presented at the beginning of your response. Only one <output_3D> is allowed, and </output_3D> should not be included.
>
> For example:
>
> <output_3D>
> <think> I have imagined a 3D scene shown above based on the given images, where there is a white water cup placed on a table. The room has distinct features on each side, including walls, windows, a fridge, and possibly furniture like cabinets and chairs. The cameras capturing the images are positioned at the front, left, back, and right of the cup, providing different perspectives of the surroundings. From the perspective of image 1 provided, we notice that there is a window with daylight coming through, and a fridge is visible next to it. This orientation suggests that the specific viewpoint from image 1 is facing these objects: the window and the fridge. In my mind, I imagine adjusting the position of the 3D scene to align it with the direction shown in image 1. Now, I need to think about what might be behind the viewpoint based on this setup. I imagine myself in the 3D scene, walking to the opposite side, that is, rotating the view horizontally 180 degrees. I saw the Window and fridge, which is what image 2 shows. So, the answer is B. Window and fridge </think>
> <answer>B. Window and fridge</answer>

Figure 5. System prompt for the CoT generation .

reasoning process o. Fig. 4 and Fig. 5 illustrate the prompts we designed for generating CoT data. Additionally, we assess the quality of the generated CoT, ensuring that o contains `<output_3D>` at the beginning, followed by the text in adherence to the format `<think>...</think> <answer>...</answer>`. The CoT data o that do not conform to this specification will be re-generated by M. If the data fails to meet the required format after more than $N = 10$ attempts, it will be discarded.

5. Explanation of the Claim

Annotation-free: In fact, unlike previous purely textual CoT methods, our method requires 3D mentaling during the reasoning process. Thus, "annotation-free" refers specifically to the absence of *manual geometric annotations for 3D mentaling* (e.g., ground-truth point clouds). In our framework, GPT-4.1 is only used to generate 3D placeholders, while the actual 3D geometric features associated with these placeholders are provided by VGGT aggregator.

VGGT Dependence: In the supervised training stage, we aim to elicit "think with 3D" behaviour without relying on manual geometric annotations, which leads us to make use of features extracted by VGGT. In contrast, learning to generate 3D geometry entirely from scratch would necessitate explicit geometric supervision. On the other hand, in the reinforced training stage, learning can be driven *solely by the outcome-based reward signal*. Specifically, as shown in Tab. 6 of the main text, even in the absence of r_{3D} there is still an **8.93%** (68.3 vs. 62.7) improvement, indicating that *3DThinker remains robust without the teacher model*.

6. Cost

Our cost reduction primarily stems from two factors: (1) *3DThinker* typically reaches near-optimal performance within 500 training steps (see Fig. 3); (2) with VLLM acceleration, RL rollout efficiency is substantially improved. Consequently, taking Qwen2.5-VL-3B as an example, on a single H200 GPU with a batch size of 1, the convergence time for supervised training and reinforcement training is **21.84 h** and **12.85 h**, respectively.

7. More Visualization

7.1. Thinking with 3D Visualization

In Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 10, Fig. 11, and Fig. 12, we provide additional visualizations on MindCube-Tiny of *3DThinker-S1+S2*$_{\text{Qwen2.5-3B}}$, including the reasoning output, 3D imagery, etc.. *The 3D imagery (reconstructed 3D point cloud) shows what the VLM is "thinking" during spatial reasoning*.

7.2. Failure Case and Limitation

As shown in Fig. 13, we present a representative failure case. Given two images from the indoor scene, *3DThinker* incorrectly infers the spatial position of the rear object. Fortunately, by visualizing the 3D mental representation, we can identify the underlying cause. The reconstructed 3D point cloud exhibits structural inconsistencies—specifically, *the wall with two photographs is mistakenly placed on the right side of the frontal window view, whereas it should appear on the left*. This misalignment likely stems from the sparsity of the two input views, which *provide insufficient spatial constraints, leading to inaccurate 3D reconstruction and, consequently, erroneous spatial reasoning*. Therefore, in the case of extremely sparse views, better integration of other information is required, which could serve as a promising direction for future work.

References

[1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. 3

[2] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. *arXiv preprint arXiv:2503.13111*, 2025. 1

[3] Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. *arXiv preprint arXiv:2509.06266*, 2025. 1, 2, 3

[4] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[5] Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. Mllms need 3d-aware representation supervision for scene understanding. *arXiv preprint arXiv:2506.01946*, 2025. 2

[6] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 2

[7] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025. 1

[8] Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025. 1

[9] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. 1

[10] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. 1

[11] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. 1

[12] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025. 1, 2, 3

[13] Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025. 1

[14] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. 2

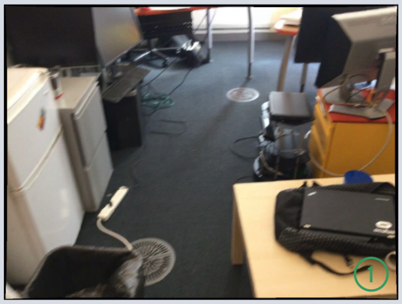
Question: Based on these three images (image 1, 2, and 3) showing the same scene from different viewpoints (back, left, and right): If I am standing at the same spot and facing the same direction as shown in image 2, then I turn left and go straight, will I get closer to the wooden table? A. It would become closer to me. B. The distance wouldn't change much, but it would be on my back-right. C. The distance wouldn't change much, but it would be on my back-left. D. It would become farther from me.

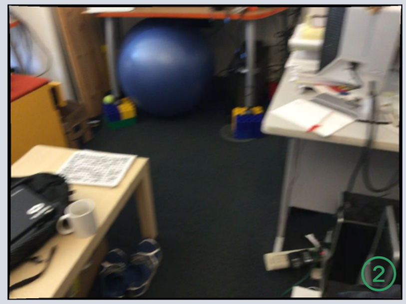
***3DThinker*:**<|latent_start|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_end|>
<think> First, I have mentally constructed the 3D layout above.
- Image 1 shows the table and chair from the back view, looking toward display cases.
- Image 2 is the left-side view, showing the chair's back and the table edge, with more of the green-covered cabinet visible.
- Image 3 is the right-side view, where bookshelves and cabinets are seen behind the table.
If I'm standing at the spot and orientation of image 2 (left side), I'm facing the table and chair, with the chair directly in front of me and the table extending to my left. The green-covered cabinet is to my left, and the open space extends ahead.
Turning left from this position means I rotate 90 degrees counterclockwise, so now I'm facing perpendicular to my previous direction—toward the open space beside the cabinets and away from the table. Going straight in this new direction means walking parallel to the long edge of the table, moving further away from it rather than approaching it.
Thus, as I walk straight after turning left, the table—which was formerly beside me—is now getting left behind me and the distance increases. Therefore, I am moving farther away from the table.
</think>
<answer>**D. It would become farther from me.**</answer>

GT: D. It would become farther from me.

Figure 6. More visualization results of our method. The reasoning process, GT answer, and the imagery of reconstructed point cloud from 3D latent are presented.

Question: These two images (image 1 and 2) show the same scene from two different viewpoints. Image 2 was taken after turning the camera 90 degrees to the right (clockwise) from the position of image 1. Based on these two images: If I am standing at the same spot and facing the same direction as shown in image 1, what is to my right? A. Blue ball B. Window





***3DThinker*:**<|latent_start|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_end|>
<think> I have constructed a 3D mental map of this room using the two images. In image 1, we see a fridge on the left side, some desks, computer monitors, and cables. There is a window in the background, centered high on the wall, providing natural light. On the right edge of image 1, the yellow cabinet and a blue ball can be seen.
In image 2, the viewpoint has rotated 90 degrees to the right (clockwise) compared to image 1. Now, most of the blue ball is visible in the center of the image, and the periphery of the room such as the table, cup, shoes, and cabinet have shifted positions relative to the ball. The right edge of this image also shows the same desk/cabinet, which links the spatial continuity to image 1.
So, if I am standing at the original spot facing the direction shown in image 1, my right side would correspond to the viewpoint seen in image 2. Since the blue ball is prominently visible in that direction after a 90-degree right turn, the correct answer for what's to my right while facing image 1's direction is A. Blue ball.
</think>
<answer>**A. Blue ball**</answer>

GT: A. Blue ball

3D Imagining

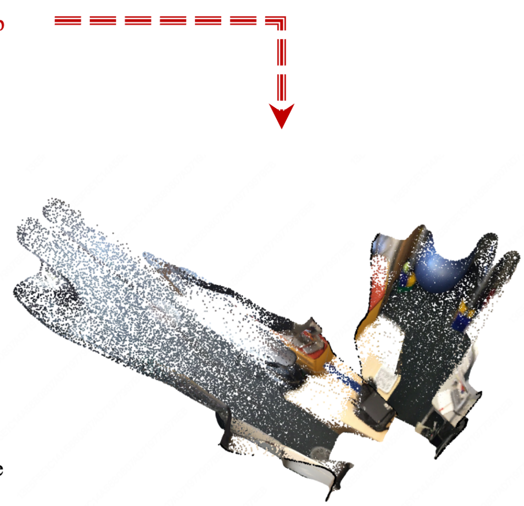


Figure 7. More visualization results of our method. The reasoning process, GT answer, and the imagery of reconstructed point cloud from 3D latent are presented.

Question: Based on these two views showing the same scene: in which direction did I move from the first view to the second view? A. Diagonally forward and left B. Directly left C. Diagonally forward and right D. Directly right

***3DThinker*:**<|latent_start|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_end|>
<think>To create the mental 3D layout, I observe that the first image shows two trash bins (Residual Waste on the left, Recyclable on the right) side-by-side, with a signboard just behind them and some trees/vegetation off to the right side and a sidewalk/hedge area to the left. The camera is facing the front of the bins directly.
In the second image, the perspective has changed: now the camera is positioned behind the bins looking toward their sides (we see the side of the Residual Waste bin and the back of the Recyclable bin). The signboard pole is now closer in view and seen from an angle, and the background details (like the road and greenery) have shifted, showing more of the adjacent pavement.
This change means the viewpoint moved from in front of the bins (Image 1) to a position further back and to the right, near the opposite corner, looking diagonally so that the bins' sides and the signboard pole are both in the foreground. The surrounding vegetation and street layout confirm this is not a simple left or right sidestep, but a movement diagonally forward and right around the bins.
Thus, the correct answer is C. Diagonally forward and right.</think>
<answer>**C. Diagonally forward and right**</answer>

GT: C. Diagonally forward and right

3D Imagining

Figure 8. More visualization results of our method. The reasoning process, GT answer, and the imagery of reconstructed point cloud from 3D latent are presented.

Question: Based on these three images (image 1, 2, and 3) showing the same scene from different viewpoints (front, left, and right): If I am standing at the same spot and facing the same direction as shown in image 3, then I turn left and go straight, what would happen to my relative position to printer? A. The distance wouldn't change much, but it would be on my front-left. B. It would become farther from me. C. The distance wouldn't change much, but it would be on my front-right. D. It would become closer to me.

***3DThinker*:**<|latent_start|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_end|>

<think>First, let's reconstruct the 3D layout using the three images.

Image 1 shows the front view of the printer and a small cabinet positioned side by side against a wall. Image 2 is taken from the left side of the printer, showing the left-side features, and we see the open space of the room beyond the printer. Image 3 is taken from the right side of the cabinet, looking left toward the side of the cabinet, with the open room visible beyond.

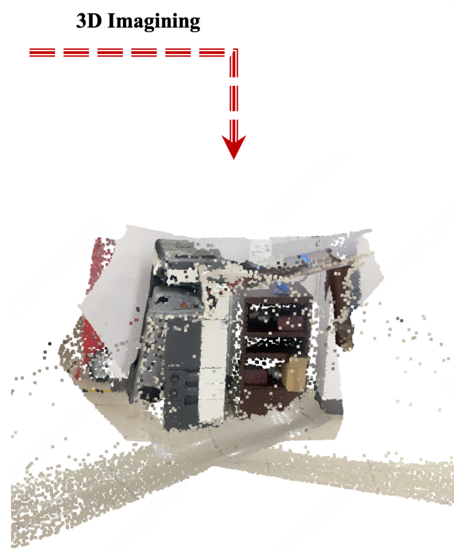
The question states that if you are standing at the same spot and facing the same direction as in image 3 (which is looking toward the side of the cabinet, with the printer and wall behind the cabinet), then you turn left and go straight.

If you are facing the direction shown in image 3, your body is oriented parallel to the wall, with the printer and cabinet on your right. Turning left means you will now face away from the wall, into the open space of the room. "Going straight" from this new orientation moves you further away from the wall (and thus from the printer, which remains near the wall).

Visualizing this: as you walk straight after turning left, you are moving perpendicularly away from the line connecting the wall and printer/cabinet. Therefore, your distance to the printer will increase.

So, the correct answer is B. It would become farther from me.</think>

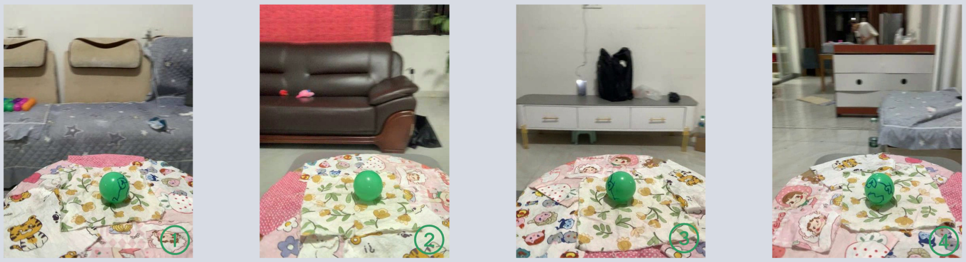
<answer>**B. It would become farther from me.**</answer>



GT: B. It would become farther from me.

Figure 9. More visualization results of our method. The reasoning process, GT answer, and the imagery of reconstructed point cloud from 3D latent are presented.

Question: Based on these four images (image 1, 2, 3, and 4) showing the green ball from different viewpoints (front, left, back, and right), with each camera aligned with room walls and partially capturing the surroundings: If I am standing at the same spot and facing the same direction as shown in image 3, then I turn right and move forward, will I get closer to the white-red cabinet? A. No B. Yes



***3DThinker*:**<|latent_start|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_end|>
<think>
To solve this, I build a mental 3D map of the room based on the four images:
- Image 1: Facing a beige sofa with colored toys on the left. This is the "front" view.
- Image 2: Facing a brown leather couch with a red curtain behind it. This is the "left" view.
- Image 3: Facing a long white-red cabinet (white body, red top). This is the "back" view.
- Image 4: Facing a white-red counter/cabinet, a bed is next to it, and you can see further into the depth of the room, including a kitchen or dining area. This is the "right" view.

All images have the green ball and cartoon-patterned cloth in the center foreground, indicating that the table with the ball is central to all shots, and the camera positions are rotated around it.
Given the question:
"If I am standing at the same spot and facing the same direction as shown in image 3, then I turn right and move forward, will I get closer to the white-red cabinet?"
Let's clarify directions:
- Facing "back" (image 3) means you're looking directly at the long white-red cabinet.
- Turning "right" from there means you now face in the direction that was called "right" in image 3, which is towards what appears to be the kitchen/counter area (as seen in image 4).
- Moving "forward" after turning right means walking in that direction—towards the kitchen/counter area.

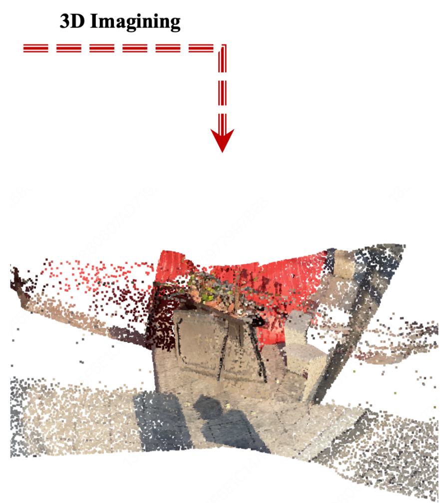
The white-red cabinet shown in image 3 is also visible next to it in image 4 (to the right side). As you turn right from facing the cabinet (image 3) and move forward (towards the side shown in image 4), you actually walk alongside the white-red cabinet, getting closer to it compared to when you started (which was directly facing the cabinet).
Therefore, moving forward after turning right does bring you nearer to the white-red cabinet/counter, matching the answer "Yes".
</think>
<answer>**B. Yes**</answer>

GT: B. Yes

Figure 10. More visualization results of our method. The reasoning process, GT answer, and the imagery of reconstructed point cloud from 3D latent are presented.

Question: Based on these three images (image 1, 2, and 3) showing the same scene from different viewpoints (front, left, and right): From the viewpoint presented in image 3, what is the nearest object behind of the stone stool. A. planter box B. water-filled barrier C. blind passage

***3DThinker*:**<|latent_start|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_end|>
<think>
I have constructed a 3D mental model of the scene based on the three provided images. There is a paved area with red temporary barriers, a square planter box containing potted plants, and a stone stool (rectangular block).
- Image 1 (front view): Shows the face of the planter box with the stone stool to the right side, both aligned near the red barriers.
- Image 2 (left view): Provides a perspective from the left, emphasizing the planter box's oblique edge and the stone stool's position directly adjacent to the box's front face.
- Image 3 (right view): Shows the scene from the right, with the stone stool in the foreground, and the planter box extending directly behind it.
The question asks: From the viewpoint in image 3 (right view), what is the nearest object behind the stone stool?
Analyzing image 3, the stone stool is closest to the camera, with the planter box extending directly behind it in the visual frame. The red barriers and pavement tiles are further back. Importantly, the planter box is positioned immediately behind the stool, whereas the other options (blind passage or water-filled barrier) are much farther away or to the sides.
Therefore, by mentally placing myself at the right view depicted in image 3, looking at the stone stool, and considering what is directly behind it relative to my perspective, I can confirm that the nearest object behind the stone stool is the planter box.
</think>
<answer>**A. planter box**</answer>



GT: A. planter box

Figure 11. More visualization results of our method. The reasoning process, GT answer, and the imagery of reconstructed point cloud from 3D latent are presented.

Question: Based on these three images (image 1, 2, and 3) showing the same scene from different viewpoints (front, left, and right): From the viewpoint presented in image 3, is there a lamp behind the laptop? image_3.png A. No. B. Yes

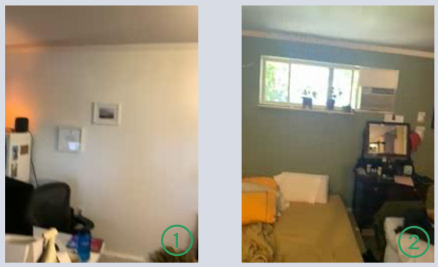
3DThinker:<|latent_start|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_end|>
<think>
I have mentally constructed the 3D layout based on the three images provided:
- Image 1 (front view): Shows a table with two wooden lamps on it. The laptop is placed between the lamps, and the setup is in a library environment with more tables and shelves visible in the background.
- Image 2 (left view): Gives a perspective from the left side of the table, revealing the placement of the first lamp closer to the camera. Behind this lamp, the second lamp is slightly further away, and the laptop is visible.
- Image 3 (right view): Focuses on the right side of the setup, with the right lamp prominently in the foreground. Looking past it, you see the laptop centrally positioned, and the left lamp is visible behind the laptop.
From the spatial arrangement in image 3, when standing on the right side of the table and looking straight across the setup, you observe that the lamp is placed directly in front of the laptop, only slightly separated by the table's edge. The left lamp, which is farther from the right side, is clearly visible behind the laptop from this perspective. This means that, from the viewpoint in image 3, there is indeed a lamp located behind the laptop.
Thus, the correct answer is: B. Yes
</think>
<answer>**B. Yes**</answer>

GT: B. Yes

3D Imagining

Figure 12. More visualization results of our method. The reasoning process, GT answer, and the imagery of reconstructed point cloud from 3D latent are presented.

Question: These two images (image 1 and 2) show the same scene from two different viewpoints. Image 2 was taken after turning the camera 90 degrees to the right (clockwise) from the position of image 1. Based on these two images: If I am standing at the same spot and facing the same direction as shown in image 2, then turn 90 degrees to the right, what is to my behind? A. Two photographs B. Window



***3DThinker*:**<|latent_start|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_pad|><|latent_end|>

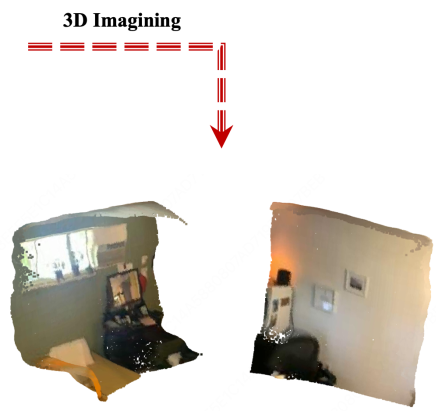
<think>I have constructed a 3D scene of a room based on the two images. In image 1, I see a wall with two photographs hanging above a chair and some other furniture, including a fridge or cupboard on the left. In image 2, after rotating the viewpoint 90 degrees to the right (clockwise), I now see a window above a desk, an air conditioner, and some curtains. There is also a bed or couch beneath the window.

Given the question, if I start by facing the direction shown in image 2 (towards the window) and then turn 90 degrees to the right, my new view will be directed towards a wall not pictured here (as the wall with photographs from image 1 would be to my right after the turn). The crucial part of the question is: after making this turn, what is now directly behind me?

If facing the window (image 2) and turning 90 degrees to the right, my back will be towards the window (since it was originally in front of me before turning). Thus, the object directly behind me will be the window, matching option B.

The two photographs (option A) would be to my right or left after the turn, not directly behind. So, the correct answer is B. Window.</think>

<answer> **B. Window.**</answer>



GT: A. Two photographs

Figure 13. Visualization of the failure case. The reasoning process, GT answer, and the imagery of reconstructed point cloud from 3D latent are presented.