PseudoGD: Enhancing Spatial Reasoning in Vision-Language Models through Pseudo Geometric Knowledge Distillation

Anonymous ACL submission

Abstract

Recent Large Vision-Language Models (LVLMs) have shown remarkable success in general semantic understanding. However, they still struggle with 3D spatial reasoning tasks, such as estimating metric distances or understanding precise relative positions. Previous works, like SpatialVLM, tried to address this by using synthesized spatial VQA dataset. However, they are fundamentally limited because their vision encoders are biased toward 2D patterns learned from image-text pairs. In this paper, we argue that this lack of 3D awareness is a critical bottleneck that cannot be solved by data scaling alone. To address this, we propose Pseudo Geometric Distillation (PseudoGD), a framework designed to help vision encoders internalize 3D geometric information using only standard 2D images. PseudoGD explicitly injects metric scale and structural context into the encoder through a Joint Training strategy. This approach optimizes geometric learning and spatial VQA tasks together, ensuring that the Large Language Model (LLM) aligns well with the improved visual features in real-time. Extensive experiments on the OmniSpatial benchmark demonstrate that PseudoGD achieves enhanced performance across various model architectures. Notably, significant improvements in Hypothetical Perspective Taking and Locate tasks prove that our model has effectively learned a physical sense of space.

1 Introduction

The scope of Vision-Language Models (VLMs) now extends beyond basic tasks like image captioning and VQA to Embodied and Physical AI, where agents interact with the real world (Radford et al., 2021; Alayrac et al., 2022; Driess et al., 2023; Li et al., 2022; Goyal et al., 2017; Zitkovich et al., 2023; Hudson and Manning, 2019). Robust spatial understanding, such as accurate perception of the locations, distances, and scales of objects, is essential for these systems (Chen et al., 2024; Wang et al., 2024; Yu et al., 2025). However, spatial reasoning remains a persistent bottleneck for current VLMs, limiting their effectiveness in tasks that require precise geometric and structural understanding (Lin et al., 2024; Liu et al., 2024; Kamath et al., 2023; Majumdar et al., 2024; Nikolov et al., 2025; Song et al., 2025).

To address this, SpatialVLM (Chen et al., 2024) improved spatial understanding by using large synthetic Spatial VQA datasets without explicit 3D training. Following this trajectory, subsequent studies have explored various avenues to increase spatial awareness. Based on this, OmniSpatial (Mengdi Jia et al., 2025) introduced a spatial reasoning benchmark grounded in cognitive psychology, VLM-3R (Fan et al., 2025) used 3D reconstruction, and SpatialRGPT (Cheng et al., 2024) improved spatial reasoning with region-based prompting.

Despite these advances, vision encoders of most VLMs are trained on 2D image-text pairs, limiting their ability to comprehend 3D spatial structures. To bridge this gap, we introduce Pseudo Geometric Knowledge Distillation (PseudoGD). This approach empowers vision encoders to internalize 3D geometric cues from monocular inputs by leveraging depth and segmentation models as "Teachers," effectively mitigating the inherent 2D bias. Unlike previous works (Huang et al., 2023; Li et al., 2025; Hong et al., 2023) that depend on 3D point clouds or multi-view data, our approach ensures high scalability and generalization performance through Pseudo knowledge distillation using only 2D images. Consequently, PseudoGD achieves a remarkable 59.6% accuracy, nearly doubling the performance of SpaceLLaVA (32.1%) and establishing a new baseline for comprehensive spatial understanding.

2 Related Works

Early VLM research primarily focused on visual grounding, establishing correspondences between visual objects and textual descriptions (Mao et al., 2016; Nagaraja et al., 2016; Yu et al., 2016). While datasets such as Visual Genome (Krishna et al., 2017) contributed to training 2D positional information at the bounding box level, they remained limited to planar recognition, excluding depth and 3D structural contexts. However, the advancement of Embodied AI and robotics has necessitated that VLMs possess 3D spatial understanding capabilities, such as metric distance estimation and relative spatial relations, beyond simple localization (Zhu et al., 2024; Sun et al., 2025).

The most dominant trend involves solutions through datasets synthesis. SpatialVLM (Chen et al., 2024) demonstrated that quantitative data expansion can enhance qualitative reasoning by constructing a massive VQA dataset that synthesizes 3D geometric information onto internet-scale 2D images. Similarly, RoboSpatial (Song et al., 2025) combined 3D scan data from robotics environments with 2D images to train the spatial awareness required for robotic manipulation. Meanwhile, SpatialRGPT (Cheng et al., 2024) and SR-3D (Cheng et al., 2025) were proposed to improve inference without modifying the model architecture. These methods enhance spatial reasoning performance by using region-based prompting to guide the model's focus toward specific pixel areas. However, these studies share a common limitation: they rely on pre-trained encoders (e.g., CLIP (Radford et al., 2021)) fundamentally biased toward 2D semantic matching. Since adjustments at the text or prompt level do not fundamentally alter the encoder's internal representations, reasoning in the absence of 3D structural information remains superficial (Hu et al., 2025; Qin et al., 2025).

Recently, model-centric approaches have emerged to geometrically tune vision encoders (Radford et al., 2021; Oquab et al., 2023; Dosovitskiy, 2020) themselves. VLM-3R (Fan et al., 2025) introduced an auxiliary module for 3D reconstruction from monocular video to assist visual perception, while 3D VLM-GD (Lee et al., 2025) proposed a geometric knowledge distillation method that extracts geometric cues from 3D foundation models and injects them into the vision encoder. Although 3D VLM-GD (Lee et al., 2025) aligns with our technical trajectory, it faces a decisive constraint stemming from its strict dependency on fine-tuning with specific datasets paired with multi-view images or 3D point clouds. The construction of such datasets necessitates specialized capture equipment and strictly controlled environments. This dependency imposes a critical bottleneck on data scalability, fundamentally contradicting the philosophy of data abundance advocated by prior works like SpatialVLM. Consequently, it structurally precludes the utilization of vast web-scale 2D data, thereby isolating the model from the rich, diverse visual distributions required for universal spatial reasoning.

3 Methodology

We introduce PseudoGD, a framework designed to empower vision encoders to internalize 3D geometric reasoning directly from monocular 2D inputs. In this section, we define the fundamental cognitive bottlenecks inherent in existing VLM training paradigms and detail how our Pseudo Geometric Distillation and Joint Training strategies effectively bridge this gap.

3.1 Pseudo Geometric Knowledge Distillation (PseudoGD)

As illustrated in Figure 1, the core principle of this technique is to transfer the geometric reasoning capabilities of two teacher models, which are Depth Pro (Bochkovskii et al., 2024) and Segment Anything Model (SAM) (Kirillov et al., 2023; Ravi et al., 2024; Carion et al., 2025), to the vision encoder without requiring explicit 3D ground-truth data. By leveraging the knowledge distillation, this method enables the encoder to internalize spatial cues that are often absent in standard vision-language pre-training.

We integrate two complementary geometric properties to enrich visual representations. First, we employ Depth Pro as the Metric Depth Teacher to inject precise metric scale information. Unlike relative depth estimation, Depth Pro accounts for focal length and physical dimensions, providing the encoder with a physical sense of scale essential for quantitative reasoning (e.g., "5-meter distance"). Second, we utilize SAM as the Structural Segmentation Teacher to instill structural context. By encapsulating sophisticated object boundaries and part-whole relationships, SAM embeddings enhance the encoder's perception of complex spa-

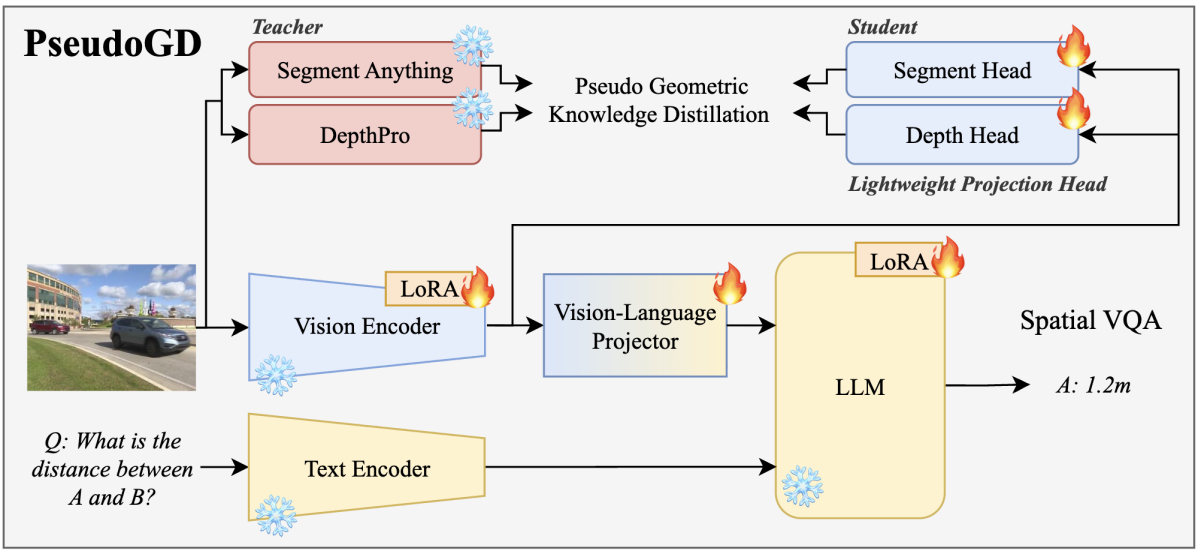

Figure 1: The training framework of PseudoGD. Our approach performs PseudoGD and Spatial VQA training simultaneously to enhance 3D spatial understanding.

tial arrangements, such as occlusions and relative positioning.

Through lightweight projection heads, we map the student encoder's features into the teachers' respective spaces. The distillation objective combines metric depth error and structural similarity loss, ensuring the model is not confined by the biases of specific 3D datasets. This facilitates the learning of universal geometric features, thereby securing a robust generalized capacity for comprehensive 3D spatial understanding.

3.2 Joint Training Strategy

Sequential training, where an encoder is pre-trained on geometric tasks before being connected to an LLM, often leads to catastrophic forgetting of semantic knowledge and feature misalignment, where the LLM fails to adapt to the shifted visual distribution. To circumvent these issues, we adopt a Joint Training strategy that co-optimizes geometric distillation and spatial VQA objectives within a unified loop. Formally, the total objective function is defined as Equation (1).

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{VQA}} + \mathcal{L}_{\text{Depth}} + \mathcal{L}_{\text{Seg}} \quad (1)$$

Where $\mathcal{L}_{\text{VQA}}$ denotes the autoregressive language modeling loss, while $\mathcal{L}_{\text{Depth}}$ and $\mathcal{L}_{\text{Seg}}$ represent the distillation losses derived from the Metric Depth and Structural Segmentation teachers, respectively.

This integrated optimization process enables the vision encoder to acquire geometric inductive biases while simultaneously allowing the LLM to learn, in real-time, how to interpret these evolving representations. Consequently, PseudoGD achieves the capability to immediately leverage visual depth and structural cues for linguistic reasoning, ensuring a seamless alignment between visual perception and language generation.

4 Experiments

4.1 Experimental Setup

Datasets. For training, we utilize the VQASynth dataset collection for SpaceLLaVA, SpaceQwen, and SpaceThinker (Chen et al., 2024), which is publicly available on Hugging Face, to establish a baseline of fundamental spatial comprehension. In the evaluation phase, we adopt the OmniSpatial benchmark (Mengdi Jia et al., 2025) as the held-out test set to rigorously validate comprehensive spatial intelligence across diverse cognitive domains.

Evaluations. We conduct a comparative evaluation against four representative VLMs that demonstrate strengths in spatial perception or general multimodal capabilities, specifically SpaceLLaVA (Chen et al., 2024; Liu et al., 2023), SpaceMantis (Chen et al., 2024), SpaceQwen2.5-VL, and SpaceThinker-Qwen2.5(Bai et al., 2023; Yang et al., 2025; Bai et al., 2025; Chen et al., 2024). These models represent the state-of-the-art in spatial interaction and general visual reasoning, thereby providing a rigorous standard.

To rigorously evaluate comprehensive spatial

Method	Overall	Dynamic Reasoning		Spatial Interaction			Complex Logic		Perspective Taking		
		Manipulate	Motion	Traffic	Locate	Geospatial Strategy	Pattern Recog.	Geometric Reasoning	Ego	Allo	Hypo.
SpaceLLaVA-13B	36.14	52.70	21.39	43.53	38.10	44.55	23.71	32.90	58.82	38.03	45.78
+ **PseudoGD**	38.88	56.76	30.35	50.59	41.90	47.27	28.87	25.81	53.92	39.10	45.78
SpaceMantis-8B	36.01	52.70	35.55	36.47	34.29	33.64	35.05	21.94	52.94	36.44	32.53
+ **PseudoGD**	37.18	54.05	33.24	42.35	35.24	36.36	34.02	21.94	51.96	38.83	43.37
SpaceQwen2.5VL-3B†	40.25	58.11	39.88	41.18	40.95	40.91	29.90	25.81	63.73	38.83	39.76
+ **PseudoGD**	39.73	50.00	42.20	48.24	43.81	37.27	28.87	27.10	50.00	36.97	45.78
SpaceThinker-Qwen2.5†	40.42	47.84	53.06	43.29	35.43	38.73	24.33	28.00	58.04	35.11	31.08
+ **PseudoGD**	42.20	55.41	49.42	49.41	40.95	39.09	26.80	26.45	67.65	35.64	44.58

Table 1: OmniSpatial benchmark results (%) on task-level evaluation. Results marked with † are cited from OmniSpatial (Mengdi Jia et al., 2025).

reasoning capabilities, we utilize the OmniSpatial evaluation set (Mengdi Jia et al., 2025) as our primary benchmark. We measure model performance across the four core dimensions defined by the benchmark: Dynamic Reasoning, Complex Spatial Logic, Spatial Interaction, and Perspective-Taking.

Implementation Details. To ensure a rigorous comparative analysis, we trained all models using the identical base architectures and datasets as their respective baselines, with the inclusion of PseudoGD being the sole experimental variable. All models were fine-tuned using LoRA (Hu et al., 2022). To balance reasoning capacity with the preservation of visual priors, we set the LoRA rank to 128 for the LLM and 4 for the vision encoder, while fully fine-tuning the multimodal projector to ensure robust cross-modal alignment. Additionally, the depth estimation and segmentation heads were implemented as lightweight modules integrated directly into the vision encoder, minimizing architectural complexity while facilitating geometric internalization. Further details are provided in the Appendix.

4.2 Experimental Results and Analysis

Superior Performance and Generalization. As shown in Table 1, PseudoGD consistently enhances spatial reasoning across diverse architectures, bridging the gap between 2D perception and 3D cognition. Notably, SpaceThinker-Qwen2.5 + PseudoGD achieved the best overall accuracy of 42.20% (+1.78%), with significant gains in Locate and Traffic Analysis. This confirms that internalizing metric depth empowers models to perform precise localization and dynamic reasoning.

Bridging Cognitive Bottlenecks. The most profound impact is observed in Hypothetical Perspective Taking, where SpaceThinker surged from 31.08% to 44.58%. This dramatic gain suggests the vision encoder has successfully internalized 3D structural information, overcoming the cognitive bottleneck of simulating unseen viewpoints without explicit 3D priors. Additionally, the universal improvement in the Locate metric proves that PseudoGD equips models with a physical sense of space, enabling reliable spatial grounding even in complex environments where semantic features alone are insufficient.

5 Conclusions

In this work, we addressed the limitations of current VLMs in 3D spatial reasoning, which stem from their reliance on 2D semantic priors. To mitigate this, we proposed PseudoGD, a framework integrating Pseudo Geometric Knowledge Distillation with a Joint Training strategy. This approach enables vision encoders to learn 3D geometric representations from monocular 2D inputs, effectively utilizing depth and segmentation cues without requiring explicit 3D training data. Our evaluation on the OmniSpatial benchmark demonstrates that this method consistently improves spatial reasoning capabilities across diverse architectures. The performance gains observed in Locate and Hypothetical Perspective Taking indicate that the model has effectively internalized physical scale and structural relationships. These findings suggest that explicitly distilling geometric features is a valid approach for enhancing the spatial understanding of VLMs.

Limitations

Although this study demonstrates that PseudoGD effectively enhances the spatial reasoning capabilities of VLMs by distilling geometric knowledge, several limitations remain.

First, the performance of our framework is fundamentally dependent on the quality of the teacher signals. Since we rely on Depth Pro (Bochkovskii et al., 2024) and SAM (Kirillov et al., 2023) to generate pseudo-labels for metric depth and structural segmentation, any errors or artifacts produced by these models inevitably propagate to the student encoder. Consequently, the model may exhibit performance degradation in scenarios where the teacher models struggle, such as scenes containing mirrors, transparent materials, or extreme lighting conditions that cause visual ambiguity.

Second, there is a limitation regarding the approximation of 3D geometry. While our method successfully empowers the vision encoder to internalize 3D cues from monocular 2D images, this remains an implicit approximation rather than an explicit measurement. Compared to approaches utilizing ground-truth 3D point clouds or multi-view geometry, our model may lack precision in fine-grained metric estimation for complex, heavily occluded structures. Future work is needed to bridge the gap between internalized 2D spatial cues and absolute 3D physical accuracy.

Third, the current framework is constrained to static single-image inference. Although the model showed improvements in the Dynamic Reasoning track of OmniSpatial, it infers motion and temporal relationships based solely on static visual evidence. Practical robotic applications often require continuous reasoning over temporal sequences to handle dynamic physical interactions. Extending the PseudoGD mechanism to video-based VLMs to capture temporal context remains a critical direction for future research.

Finally, the computational overhead during training is non-negligible. Unlike standard instruction tuning that relies on sparse token prediction, our joint training strategy involves aligning dense, pixel-level features from projection heads with teacher embeddings. This increases memory consumption and computational cost during the training phase, presenting a challenge for scalability when applying this method to ultra-large-scale multimodal models.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Aleksei Bochkovskii, AmaÃĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. 2024. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

An-Chieh Cheng, Yang Fu, Yukang Chen, Zhijian Liu, Xiaolong Li, Subhashree Radhakrishnan, Song Han, Yao Lu, Jan Kautz, Pavlo Molchanov, et al. 2025. 3d aware region prompted vision language model. *arXiv preprint arXiv:2509.13317*.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. 2025. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2023. 3d concept learning and reasoning from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9202–9212.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. 2025. G^{2}vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. *arXiv preprint arXiv:2511.21688*.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2023. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Seonho Lee, Jiho Choi, Inha Kang, Jiwook Kim, Junsung Park, and Hyunjung Shim. 2025. 3d-aware vision-language models fine-tuning with geometric distillation. *arXiv preprint arXiv:2506.09883*.

Chengmeng Li, Junjie Wen, Yan Peng, Yaxin Peng, Feifei Feng, and Yichen Zhu. 2025. Pointvla: Injecting the 3d world into vision-language-action models. *arXiv preprint arXiv:2503.07511*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. 2024. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Shaochen Zhang Wenyao Zhang Xinqiang Yu Jiawei He He Wang Li Yi Mengdi Jia, Zekun Qi et al. 2025. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *Submitted to The Fourteenth International Conference on Learning Representations*. Under review.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer.

Nikolay Nikolov, Giuliano Albanese, Sombit Dey, Aleksandar Yanev, Luc Van Gool, Jan-Nico Zaech, and Danda Pani Paudel. 2025. Spear-1: Scaling beyond robot demonstrations via 3d understanding. *arXiv preprint arXiv:2511.17411*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and Xudong Wang. 2025. Chain-of-visual-thought:

Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. 2025. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15768–15780.

Lin Sun, Bin Xie, Yingfei Liu, Hao Shi, Tiancai Wang, and Jiale Cao. 2025. Geovla: Empowering 3d representations in vision-language-action models. *arXiv preprint arXiv:2508.09071*.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. 2024. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, et al. 2025. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905*.

Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, and Tong He. 2024. Spa: 3d spatial-awareness enables effective embodied representation. *arXiv preprint arXiv:2410.08208*.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

Appendix A Implementation Details

A.1 Common Configuration

LoRA Configuration. We apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to both the language model and the vision encoder, while fully fine-tuning the multimodal projector. All LoRA modules use a dropout rate of 0.05. The detailed configuration is summarized in Table 2.

Component	Rank	Alpha	LR
LLM	128	256	2×10^{-5}
Vision Encoder	4	8	1×10^{-5}
MM Projector	Full	–	2×10^{-5}

Table 2: LoRA configuration for all experiments. LLM LoRA targets all attention projections and feed-forward layers. Vision encoder LoRA targets the fused QKV projection and output projection in each of the 32 transformer blocks.

Geometric Distillation Heads. To facilitate geometric knowledge distillation, we attach lightweight prediction heads directly to the vision encoder's output features. The depth prediction head is implemented as a Multi-Layer Perceptron (MLP) that linearly projects the input features to a 1024-dimensional hidden layer with GELU activation, followed by a final linear projection to a scalar output. To enforce physically meaningful metric constraints, we apply a scaled activation function, defined as $\text{Softplus}(x) \times 7.0 + 0.1$, ensuring positive depth values within the valid range of $[0.1, \infty)$ meters. The resulting predictions are interpolated to match the teacher's 24×24 resolution.

The segmentation head is composed of a feature projector and a spatial upsampling module designed to reconstruct high-resolution structural details. The projector utilizes an MLP ($\rightarrow 2048 \rightarrow 256$) with GELU activation to compress semantic features into a structural embedding space. Subsequently, these features are spatially reshaped and processed by Transposed Convolution layers followed by a GELU activation and a standard Convolution layer (kernel size 3, padding 1), ultimately yielding 64×64 feature maps. Both heads are optimized with a learning rate of 1×10^{-3}, significantly higher than the base model parameters, to facilitate the rapid adaptation of these randomly initialized components.

A.2 SpaceQwen2.5 w/ PseudoGD Configuration

Base Model and Dataset. SpaceQwen2.5 w/ PseudoGD is built upon `Qwen/Qwen2.5-VL-3B-Instruct`. The model is fine-tuned using the `remyxai/OpenSpaces` dataset, a spatial reasoning benchmark containing approximately 10K question-answering samples. Each sample consists of an RGB image paired with a natural language question about spatial relationships (e.g., relative positions, distances, orientations) and a corresponding answer. The dataset covers diverse indoor and outdoor scenes with varying complexity levels. The dataset is partitioned into 9.26K training, and 1.03K test samples. All experiments are conducted using `bfloat16` mixed-precision training for memory efficiency while maintaining numerical stability.

Training Hyperparameters. Training employs a batch size of 4 with gradient accumulation steps of 8, yielding an effective batch size of 32. The maximum sequence length is set to 2048 tokens to accommodate both visual tokens (variable length due to dynamic resolution) and text tokens. We employ the AdamW optimizer with no weight decay. The learning rate follows a cosine annealing schedule with 3% linear warmup. Training proceeds for approximately 3 epochs over the full dataset.

Input Processing. Images are processed at their native resolution by Qwen2.5-VL's dynamic resolution mechanism, preserving fine-grained spatial details. For teacher model inference, images are center-cropped and resized to 224×224 pixels, then normalized using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$) to ensure compatibility with the pre-trained teacher models.

A.3 SpaceThinker w/ PseudoGD Configuration

Base Model and Dataset. SpaceThinker w/ PseudoGD is based on `UCSC-VLAA/VLAA-Thinker-Qwen2.5VL-3B`, a variant of Qwen2.5-VL fine-tuned for explicit chain-of-thought reasoning. Training is performed on the `remyxai/SpaceThinker` dataset, which contains approximately 12K samples with structured reasoning annotations. The dataset is partitioned into 11.4K training, and 1.25K test samples. The model follows a two-stage output format with explicit reasoning:

```
<think>[step-by-step reasoning]</think>
<answer>[final answer]</answer>
```

This format encourages the model to externalize its spatial reasoning process before providing answers. The system prompt instructs: "*You should first think about the reasoning process and then provide the answer. Use <think>...</think> and <answer>...</answer> tags.*" All experiments use `bfloat16` precision.

Training Hyperparameters. Due to the longer output sequences required for explicit reasoning, SpaceThinker is trained with a reduced per-GPU batch size of 1 and gradient accumulation over 8 steps, yielding an effective batch size of 8. The maximum sequence length remains 2048 tokens with a warmup ratio of 3%. The model is trained for 3 full epochs over the dataset.

A.4 SpaceLLaVA-13B w/ PseudoGD Configuration

Base Model and Dataset. SpaceLLaVA-13B w/ PseudoGD is constructed upon the LLaVA-v1.5-13B architecture, utilizing CLIP ViT-L/14 as the vision encoder which produces 1024-dimensional visual features. To validate the generalizability of our geometric distillation approach, the model is fine-tuned on the `remyxai/vqasynth_spacellava` dataset, comprising approximately 28K synthetic spatial QA samples. The dataset is partitioned into 25.2K training, and 2.8K test samples.

Training Hyperparameters. Training employs a batch size of 4 with gradient accumulation steps of 8, yielding an effective batch size of 32. The model is trained for 1 epoch using AdamW optimizer and cosine annealing schedule. Mixed precision training (bfloat16) is employed for memory efficiency.

A.5 SpaceMantis w/ PseudoGD Configuration

Base Model and Architecture. SpaceMantis w/ PseudoGD extends the Mantis-8B-siglip-llama3 architecture, which integrates a SigLIP vision encoder with the Llama-3-8B language model. This configuration serves to evaluate the efficacy of geometric distillation on a multi-image capable VLM framework underpinned by a distinct vision backbone. The model is fine-tuned on the `remyxai/vqasynth_spacellava` dataset. The dataset is partitioned into 25.2K training, and 2.8K test samples.

Training Hyperparameters. The model is trained with a batch size of 4 and gradient accumulation steps of 8, resulting in an effective batch size of 32. The optimizer is AdamW and cosine annealing schedule is applied. Mixed precision training (bfloat16) is employed for memory efficiency.

A.6 Loss Function

The total training objective combines three complementary loss terms as defined in Equation (2):

$$\mathcal{L}_{\text{total}} = \lambda_{\text{vqa}}\mathcal{L}_{\text{VQA}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} \quad (2)$$

where $\mathcal{L}_{\text{VQA}}$ is the standard cross-entropy loss for next-token prediction, and $\mathcal{L}_{\text{depth}}$, $\mathcal{L}_{\text{seg}}$ are the geometric distillation losses described below. All balancing coefficients are set to $\lambda_{\text{vqa}} = \lambda_{\text{depth}} = \lambda_{\text{seg}} = 1.0$ by default, ensuring equal importance between semantic reasoning and geometric internalization.

Depth Distillation Loss. We adopt a scale-invariant logarithmic loss, which is robust to global scale ambiguities and focuses on relative depth relationships:

$$\mathcal{L}_{\text{depth}} = \frac{1}{n}\sum_i d_i^2 - \frac{1}{2n^2}\left(\sum_i d_i\right)^2, \quad (3)$$
$$d_i = \log(\hat{y}_i) - \log(y_i)$$

Here, $\hat{y}_i$ and y_i denote the predicted and teacher depth values at spatial location i, respectively, and n is the total number of spatial locations. The second term penalizes systematic scale shifts, encouraging the model to learn relative depth orderings even when absolute scale information is noisy. Depth values are clamped to a minimum of 0.1 meters before taking the logarithm to ensure numerical stability.

Segmentation Distillation Loss. The segmentation distillation loss measures the alignment between predicted and teacher feature embeddings using cosine similarity:

$$\mathcal{L}_{\text{seg}} = 1 - \frac{1}{HW}\sum_{h,w} \cos(\hat{\mathbf{f}}_{h,w}, \mathbf{f}^{\text{teacher}}_{h,w}) \quad (4)$$

where $\hat{\mathbf{f}}_{h,w}, \mathbf{f}^{\text{teacher}}_{h,w} \in \mathbb{R}^{256}$ are the L_2-normalized feature vectors at spatial position (h, w). This loss encourages the student to learn semantically meaningful spatial representations that capture object boundaries and structural patterns, without requiring explicit segmentation labels.

A.7 Teacher Models

To empower the vision encoder with robust geometric inductive biases, we employ two complementary foundation models as teachers. First, for Metric Depth Distillation, we utilize Depth Pro (Bochkovskii et al., 2024), developed by Apple. Unlike conventional relative depth estimators, Depth Pro predicts absolute metric depth (in meters), enabling the student model to internalize a physical sense of scale. For compatibility with our projection architecture, the extracted depth maps are interpolated to a 24×24 spatial resolution. Second, for Structural Segmentation Distillation, we adopt the Segment Anything Model (SAM) (Kirillov et al., 2023) with a ViT-Base backbone. SAM provides semantic-agnostic structural embeddings that encode precise object boundaries. These 256-dimensional embeddings are similarly interpolated to a 64×64 resolution, ensuring that the encoder learns to capture fine-grained structural details within a standardized feature space.

A.8 Hardware and Software

All experiments are conducted on a cluster of 5× NVIDIA A100 80GB PCIe GPUs. Multi-GPU training is orchestrated using the Hugging Face Accelerate library with distributed data parallelism. The framework is implemented in PyTorch 2.0+, leveraging the Transformers library for model loading and the PEFT library for parameter-efficient fine-tuning.

A.9 Ablation Study on PseudoGD Weight.

Table 3 presents an ablation study on the loss weighting factor λ used for PseudoGD when applied to the SpaceMantis-8B model. The loss weight λ controls the relative contribution of geometric distillation during training, allowing us to analyze how strongly enforcing geometric supervision affects different categories of spatial reasoning.

Overall, increasing the PseudoGD weight (λ) consistently improves the aggregate performance, with the best overall accuracy achieved at $\lambda = 1.0$. This trend suggests that stronger geometric distillation provides more stable and globally beneficial supervision for spatial reasoning tasks. In particular, tasks under *Spatial Interaction* and *Complex Logic*, such as *Traffic*, *Geospatial Strategy*, and *Pattern Recognition*, show noticeable gains as the weight increases, indicating that these tasks benefit

Method	Overall	Dynamic Reasoning		Spatial Interaction			Complex Logic		Perspective Taking		
		Manipulate	Motion	Traffic	Locate	Geospatial Strategy	Pattern Recog.	Geometric Reasoning	Ego	Allo	Hypo.
SpaceMantis-8B	36.01	52.70	35.55	36.47	34.29	33.64	35.05	21.94	52.94	36.44	32.53
+ **PseudoGD** (λ=0.1)	36.73	55.41	32.95	36.47	35.24	33.64	28.87	23.87	56.86	40.43	33.73
+ **PseudoGD** (λ=0.5)	36.96	55.41	33.53	38.82	35.24	35.45	31.96	23.23	54.90	39.36	38.55
+ **PseudoGD** (λ=1.0)	37.18	54.05	33.24	42.35	35.24	36.36	34.02	21.94	51.96	38.83	43.37

Table 3: OmniSpatial benchmark results (%) for SpaceMantis variants with different PseudoGD lambdas. λ contains only λ_{depth} and λ_{seg}, and the same hyperparameter values were set for both lambdas. λ_{vqa} was always fixed at 1.

from explicit geometric constraints and relational structure.

In contrast, lower PseudoGD weights (e.g., $\lambda = 0.1$) tend to favor *Perspective Taking* subtasks, especially *Egocentric* and *Allocentric* reasoning. This behavior suggests that weaker geometric regularization allows the model to rely more on implicit visual cues and viewpoint-dependent reasoning, rather than enforcing rigid geometric consistency. However, this comes at the cost of degraded performance on geometry-intensive subtasks, such as *Pattern Recognition*.

The intermediate setting ($\lambda = 0.5$) exhibits a balanced trade-off between these two regimes, achieving moderate improvements across most task categories without strongly biasing the model toward either perspective-centric or geometry-centric reasoning. These results indicate that the PseudoGD weight serves as an effective control knob for balancing geometric consistency and viewpoint flexibility in multi-modal spatial reasoning models.