G^2VLM: Geometry Grounded Vision Language Model with Unified 3D Reconstruction and Spatial Reasoning

Supplementary Material

A. Architecture Details

We employ a pretrained DINOv2 encoder. Unlike VGGT [14] or π^3 [17], where DINO features directly match the geometry transformer's hidden dimension, our architecture adopt an additional feature alignment step. Because we initialize the geometric perception expert to match the dimensions of the underlying LLM (Qwen2-VL-2B), we use a linear projection layer to map the DINO features into the expert's input space. This approach is very commonly adopted in VLM literature. Both the geometric expert and sematic expert comprises 28 layers, which mirroring the Qwen2-VL-2B architecture, and utilizes global attention.

In contrast to VGGT, which relies on a computationally intensive DPT head that aggregates multi-scale features, we adopt lightweight transformer-based geometry heads which is similar to π^3. The decoders for camera poses, local point maps, and global point maps, share the same trasformer architecture but do not share weights. This architecture is a lightweight, 5-layer transformer that applies self-attention exclusively to the features of each individual image. Following the decoder, the output heads vary by task. The heads for local point maps consist of a simple MLP followed by a pixel shuffle operation. For camera poses, the head is adapted from Reloc3r [7] and Pi3 [17] and uses an MLP, average pooling, and another MLP. The rotation is initially predicted in a 9D representation [9] and is then converted to a 3×3 rotation matrix via SVD orthogonalization. As we mentioned in Sec 3.1, the global point head serves solely to stabilize training and is excluded during inference.

B. Training Details

For the pre-training stage of geometric perception expert, we further divide it into two stage training. We first fix image resolution to 224x224 and use AdamW optimizer for 100K iterations with a learning rate (lr) of 2e-4 using cosine scheduler. Then we further train at lr of 5e-4 for another 20K steps with a higher resolution of 518x518, and apply randomized aspect ratio between 0.5 and 1.0. Similar to VGGT, for every batch, we randomly sample 2–24 frames from a random training scene. For our visual geometry loss function, we set the weights for each component as follows: $\lambda_{\text{normal}} = 1.0$, $\lambda_{\text{cam}} = 0.2$, and $\lambda_{\text{trans}} = 200.0$. The implementation of our normal loss follows that of MoGe [16] and π^3 [17], and the resolution for aligning the local point map loss is set to 4096. The low-resolution pretraining runs on 32 A800 GPUs over 7 days and high-resolution runs on 64 A800 GPUs over 3 days. For visual geometry learning, we clip the loss that is greater than 10 and smooth it to 0 to avoid training instabilities. The loss spikes are due to noisy large 3D annotation data and further data cleaning efforts can help minimize these phenomena.

For joint-training, we use AdamW optimizer and cosine scheduler for 16K iterations with a lr of 2e-5 on 64 A800 GPUs over 3 days. We do not apply loss clipping here. Throughout all training, we employ gradient norm clipping with a threshold of 1.0 to ensure training stability and leverage bfloat16 precision and gradient checkpointing to improve GPU memory and computational efficiency.

C. More results

We evaluate our model on the SPAR-Bench which is a comprehensive spatial reasoning benchmark. Here we present the detailed results of each sub-task in SPAR-Bench in Table 1. Our model, G^2VLM-SR, demonstrates the best performance consistently across all tasks. Notably, it surpasses human performance in the low category.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Anthropic. Claude 3.7 sonnet system card. https://www.anthropic.com/news/claude-3-7-sonnet, 2025. Hybrid reasoning multimodal model released November 2025. 2

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[4] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 2

[5] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v?

Methods	Rank	Avg.	Low	Depth-OC	Depth-OC-MV	Depth-OO	Depth-OO-MV	Dist-OC	Dist-OC-MV	Dist-OO	Dist-OO-MV	Medium	PosMatch	CamMotion	ViewChgI	High	DistI-OO	DistI-OO-MV	ObjRel-OC-MV	ObjRel-OO	ObjRel-OO-MV	SpImg-OC	SpImg-OC-MV	SpImg-OO	SpImg-OO-MV
Baseline																									
Chance Level (Random)	-	-	-	-	-	-	-	-	-	-	-	-	22.65	24.50	-	25.09	23.82	22.02	31.25	25.27	22.16	25.81	24.42	24.17	26.89
Chance Level (Frequency)	-	32.74	31.19	43.09	43.51	17.38	13.05	41.90	30.99	27.40	32.17	38.25	29.01	26.75	59.00	32.29	52.94	50.60	28.25	26.92	26.59	26.34	26.74	26.49	25.77
SPAR-Bench *(tiny)* API																									
Human Level	1	67.27	55.31	72.75	74.25	28.75	36.25	78.25	52.25	66.5	33.50	72.32	92	64	60.97	76.22	80	94	70	92	80	78	82	50	60
GPT-4o [1]	3	36.39	29.25	53.80	45.00	15.00	13.60	37.40	34.40	23.40	24.40	24.93	30	16	28.80	45.11	64	64	58	46	46	32	44	30	22
Claude-3.7-Sonnet [2]	5	21.77	25.43	41.00	45.40	11.20	12.20	42.60	19.60	26.00	5.40	7.33	16	6	0.00	23.33	40	48	22	36	14	12	20	6	12
Qwen2-VL-72B [15]	4	35.62	35.28	45.40	49.80	13.80	10.00	54.60	49.40	36.80	22.40	23.39	42	18	10.16	40.00	60	68	50	38	44	18	28	18	36
Qwen2.5-VL-72B [3]	2	39.40	35.35	53.20	46.80	17.80	29.00	49.60	57.40	14.40	14.60	23.05	40	16	13.16	48.44	74	74	60	56	50	20	34	24	44
SPAR-Bench *(full)*																									
LLaVA-Video-7B	5	32.33	23.55	26	37.4	26.9	12	16.4	16.3	26.9	26.5	24.83	35.4	30.8	8.3	42.62	58.2	54.2	53.8	43.1	38.8	34.9	36.6	26.5	37.5
InternVL2-2B [5]	15	28.06	21.74	18.06	24.81	23.20	20.97	19.47	19.95	26.83	20.61	22.83	39.69	23.00	5.81	35.42	51.18	55.95	46.00	31.59	23.82	36.02	34.30	17.55	22.41
InternVL2-4B [5]	7	32.01	28.94	23.94	27.22	20.00	18.12	42.57	40.16	31.29	28.18	29.16	49.87	21.00	16.62	35.70	56.76	55.36	40.25	36.81	25.21	28.76	32.27	21.19	24.65
InternVL2.5-2B [4]	11	30.14	25.79	39.67	39.72	12.12	15.03	30.94	29.59	20.22	19.02	22.93	37.91	24.25	6.64	36.41	51.47	56.85	50.25	33.79	24.10	27.15	35.17	26.49	22.41
InternVL2.5-4B [4]	10	30.55	25.66	29.06	32.97	21.77	16.83	20.84	26.85	28.13	28.79	29.75	47.07	33.25	8.92	35.16	54.12	58.93	35.50	29.67	34.63	24.73	31.39	19.21	28.29
InternVL2.5-8B [4]	3	36.28	29.46	25.78	29.31	23.79	18.76	46.82	42.68	22.62	25.89	31.88	61.32	28.00	6.32	43.80	59.71	56.85	51.75	44.23	41.55	36.56	41.57	22.52	39.50
LLaVA-OV-0.5B [10]	12	29.48	30.14	49.22	42.72	18.04	14.92	31.48	25.67	28.98	30.10	15.89	24.43	21.75	1.50	33.42	50.88	50.00	32.00	27.75	26.04	30.91	34.01	24.50	24.65
LLaVA-OV-7B [10]	8	31.20	21.79	30.33	26.94	18.58	13.87	10.43	13.64	31.24	29.29	26.13	38.68	30.25	9.47	40.14	56.47	55.06	37.25	48.63	38.23	30.38	33.72	26.49	35.01
Qwen2-VL-2b [15]	16	24.60	19.43	38.03	40.63	18.84	14.09	7.81	7.07	17.82	11.14	27.55	26.21	25.25	31.20	28.22	54.12	49.11	21.75	25.27	12.47	23.92	27.62	24.83	14.85
Qwen2-VL-7b [15]	9	30.74	27.52	35.97	35.22	20.83	12.88	28.68	29.95	28.21	28.45	20.44	35.37	20.25	5.69	37.03	59.71	52.38	30.25	38.46	41.00	22.04	28.49	22.52	38.38
Qwen2.5-VL-3B [3]	13	29.39	26.69	31.7	34.2	32.1	17.5	18.4	22.7	32.1	24.8	24.87	39.2	27.3	8.1	33.29	55.6	60.7	37.5	32.1	20.2	21	27	20.9	24.6
Qwen2.5-VL-7b [3]	4	33.07	28.75	31.33	33.66	21.99	14.97	42.88	37.73	23.83	23.64	22.97	33.33	28.75	6.83	40.27	58.24	51.49	44.75	50.00	32.13	33.87	32.85	27.15	31.93
LLaVA-v1.5-7b [11]	18	23.65	10.85	5.17	12.53	17.37	11.34	7.25	5.26	18.73	9.12	26.50	24.43	26.75	28.31	34.09	51.18	52.38	34.25	24.18	26.87	34.68	29.94	22.52	30.81
LLaVA-v1.6-7b [12]	19	13.21	8.53	12.14	0.00	20.35	0.27	10.76	0.41	24.27	0.00	4.79	6.62	7.75	0.00	20.18	51.76	7.74	6.25	32.14	6.37	39.52	10.47	21.52	5.88
SpaceMantis-13B [13]	14	28.93	23.56	35.2	29.1	18.1	13.3	21.4	23.1	24.9	23.4	23.27	31.8	31.8	6.2	35.60	56.5	55.4	41.8	31	36.3	25.3	25.6	25.5	23
SpaceQwen2.5-VL-3B [6]	17	24.24	14.46	23.7	25.8	18.9	19.1	4.7	3.3	13.1	7.1	24.00	34.4	29.5	8.1	33.01	58.5	58.6	35.5	23.9	24.1	22.6	27.9	19.9	26.1
Spatial-MLLM-7B [18]	6	32.15	29.88	31.9	22.9	22.8	16.4	35.9	38.7	35.5	34.9	20.30	34.1	26.8	0	38.13	54.7	50.9	39	34.6	24.7	38.7	41.3	28.8	30.5
VLM3R-7B [8]	2	43.21	39.78	47.8	45.6	40.1	20.6	42.2	44.3	40.1	37.5	28.43	42	30	13.3	51.18	55.9	59.2	58.8	53	54.6	47.3	50.6	30.5	50.7
G^2VLM-SR	1	54.87	59.99	80.27	73.75	21.42	18.87	78.44	75.17	68.44	63.55	36.27	27	28.2	53.3	56.51	53.5	49.1	76.8	50	68.7	50.5	52.6	44.4	63

Table 1. **Performance of different models on SPAR-Bench** The highest, second-highest, and third-highest scores in each category are highlighted with light red, light orange, and light yellow, respectively. SPAR-Bench (*tiny*) refers to a subset of the full benchmark, where 50 questions are sampled per task. Our model, G^2VLM-SR, demonstrate the best performance consistently across all tasks. Notably, it surpasses human performance in low category.

closing the gap to commercial multimodal models with open-source suites, 2024. 2

[6] RemyxAI (community) and QwenLM Team. Spaceqwen2.5-vl-3b-instruct. https://huggingface.co/remyxai/SpaceQwen2.5-VL-3B-Instruct, 2025. Fine-tuned version of Qwen2.5-VL on spatial reasoning data. 2

[7] Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16739–16752, 2025. 1

[8] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025. 2

[9] Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An analysis of svd for deep rotation estimation. In *Advances in Neural Information Processing Systems*, pages 22554–22565, 2020. 1

[10] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2024. 2

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 2

[12] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[13] remyx.ai. Spacemantis-13b. https://huggingface.co/remyxai/SpaceMantis, 2025. Vision-language model specialised for spatial reasoning; model size 13B parameters. Accessed: 2025-11-13. 2

[14] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1

[15] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[16] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking

accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5261–5271, 2025. 1

[17] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 1

[18] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 2