EgoReasoner: Metric Spatial Reasoning in VLMs for Embodied Agents

Anonymous CVPR submission

Paper ID *****

Abstract

*Embodied agents require metric spatial self-localization to act effectively, yet vision-language models (VLMs)—increasingly used as foundation models for such agents—consistently struggle with metric camera geometry. We present **EgoReasoner**, a two-stage framework that trains VLMs to estimate camera pose trajectories via geometric chain-of-thought reasoning, combining supervised fine-tuning with GRPO reinforcement learning on geometry-grounded rewards. Trained on 684,455 frames from RealEstate10K, our model achieves 3.2× and 1.7× higher structured output parse rates over a base 8B VLM and Gemini 3.1 Pro, and surpasses Gemini in translation accuracy by over 6×. These results demonstrate that structured training can instill metric geometric reasoning in VLMs, advancing them toward spatially-aware foundation models for embodied agents.*

1. Introduction

Embodied agents must localize themselves spatially to act purposefully—a prerequisite for long-horizon interaction, navigation, and object manipulation in the physical world. Vision-language models (VLMs) have become a dominant backbone for embodied agent perception [1, 10], yet they consistently fail at metric camera geometry: existing VLMs cannot reliably infer rotation magnitudes, translation scales, or trajectory consistency from image sequences. Traditional pose estimation pipelines [11, 13, 14] fill this role in isolation, but remain specialized modules disconnected from the semantic reasoning layers central to modern agent systems. Closing this gap—enabling a single VLM to jointly reason about scene semantics *and* its own metric motion—is a key step toward foundation models for spatially-aware embodied agents.

We present **EgoReasoner**, to our knowledge the first framework to train a VLM to estimate metric camera pose trajectories as an explicit reasoning task. EgoReasoner employs a two-stage pipeline: (1) supervised fine-tuning (SFT) on structured five-step geometric chain-of-thought chains that decompose pose estimation into epipolar reasoning, depth estimation, 3D triangulation, and relative pose integration; followed by (2) GRPO [12] reinforcement learning with geometry-grounded reward signals to further refine metric accuracy without degrading output format reliability. Trained on 684,455 frames from two non-overlapping RealEstate10K splits, the pipeline yields consistent, parseable trajectory outputs and strong pose accuracy across all evaluation metrics.

On a 200-sequence holdout, EgoReasoner (SFT+RL) achieves 3.2× higher structured output parse reliability than an unmodified 8B VLM and 1.7× higher than Gemini 3.1 Pro [5], while surpassing Gemini in translation accuracy by over 6×. These results demonstrate that metric geometric reasoning is learnable within VLMs through structured training, with direct implications for building foundation models capable of unified spatial and semantic perception for embodied agents.

Contributions:

- **Novel direction.** To our knowledge, EgoReasoner is the first work to formulate metric camera pose trajectory estimation as an explicit VLM reasoning task, bridging language-grounded perception and metric spatial self-localization for embodied agents.
- **Method.** A two-stage SFT+RL training pipeline combining five-step geometric chain-of-thought supervision with GRPO geometry-grounded rewards, enabling reliable structured output and progressive metric accuracy improvement.
- **Results.** EgoReasoner surpasses Gemini 3.1 Pro in translation accuracy by over 6× (RTE median 0.028 vs. 0.190) and achieves 3.2× and 1.7× higher output parse reliability than a base 8B VLM and Gemini 3.1 Pro respectively, with RRE median 1.21° on the full 200-sequence evaluation set.

2. Related Work

Camera Pose Estimation and 3D Reconstruction. Structure-from-Motion systems such as COLMAP [11] recover camera poses and sparse 3D structure from fea-

ture correspondences via bundle adjustment. Recent feed-forward methods including DUSt3R [14], MASt3R [9], VGGT [13], and π^3 [15] estimate camera geometry directly in a single forward pass, providing strong learned priors for multi-view reconstruction. However, these methods treat pose estimation as direct geometric regression without interpretable reasoning. EgoReasoner instead studies metric camera trajectory estimation as a language-mediated geometric reasoning problem within a VLM.

Vision-Language Models for Spatial Reasoning. Large VLMs such as LLaVA [10], InternVL [4], and Qwen-VL [1] have shown strong capabilities in visual question answering, captioning, and general visual reasoning. Recent works extend VLMs toward 3D spatial understanding, including SpatialVLM [3] for spatial relation reasoning and 3D-LLM [7] for 3D scene understanding. EgoReasoner targets the less-explored setting of metric camera pose trajectory inference through explicit step-by-step geometric reasoning.

Reasoning via Supervised and Reinforcement Learning. Chain-of-thought prompting [16] and RL with verifiable rewards (DeepSeek-R1 [6]) show that structured supervision and ground-truth signals can elicit complex reasoning in LLMs beyond what fine-tuning alone achieves. EgoReasoner adapts this to geometric reasoning: ground-truth camera poses serve as a natural verifiable reward, enabling VLMs to improve both final trajectory accuracy and the consistency of intermediate geometric reasoning steps.

3. Method

3.1. Problem Formulation

Given a sequence of N consecutive RGB images $\{I_1, I_2, \ldots, I_N\}$ captured by a moving camera, EgoReasoner aims to predict the camera pose trajectory $\{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_N\}$, where each $\mathbf{T}_i \in \mathrm{SE}(3)$ represents the camera-to-world transformation at frame i. Following standard practice, poses are represented as rotation matrices $\mathbf{R}_i \in \mathrm{SO}(3)$ and translation vectors $\mathbf{t}_i \in \mathbb{R}^3$.

Unlike conventional regression-based approaches, EgoReasoner formulates this as a *reasoning task*: the VLM is asked to produce a structured natural language chain of thought that explains *how* it arrives at the pose estimates, before outputting the final numerical trajectory. This formulation enables the model to leverage its pre-trained visual and language understanding to decompose the problem into interpretable sub-steps.

3.2. Geometric Reasoning Chain Construction

A key design principle is that *all mathematical steps are computed by a deterministic program*, while the VLM supplies logical inter-step reasoning insights and linguistically diverse natural language descriptions—populating a varied prompt pool and preventing the training data from degenerating into formulaic concatenations of numerical outputs.

Synthetic correspondence generation. Rather than relying on a feature matcher, we generate pixel correspondences purely by geometric projection using GT poses and depth maps from Depth Anything V3 [17] (applied jointly to 5 frames). For each sequence, 5 reference points $\{(u_j, v_j)\}$ are sampled in Frame 0. Each point is back-projected to 3D as $\mathbf{X}_j = d_j \cdot \mathbf{K}^{-1}[u_j, v_j, 1]^\top$, transformed to Frame i via the GT relative pose: $\mathbf{X}'_j = \mathbf{R}\mathbf{X}_j + \mathbf{t}$, and re-projected to get the corresponding pixel $[u'_j, v'_j] = \mathbf{K}(\mathbf{X}'_j / X'_{j,z})$ along with its depth. This yields exact 2D–3D correspondences across all frames with no feature matching required.

Chain-of-thought assembly. Each training sample consists of a user query (5 RGB frames + $\mathbf{K}$ + instruction) paired with a five-step assistant response: **(1)** point selection and matching description—*VLM-generated* natural language; **(2)** depth estimation at selected points—*VLM-generated*; **(3)** 3D back-projection—*program-filled* exact arithmetic; **(4)** Procrustes pose estimation via SVD on 3D correspondences to solve $(\mathbf{R}, \mathbf{t})$—*program-filled*; **(5)** trajectory summary—*VLM-generated*. The final numerical pose sequence is delimited by regex-parseable tokens, enabling automatic evaluation of both intermediate step quality and end-to-end trajectory accuracy.

3.3. Supervised Fine-Tuning

EgoReasoner is initialized from Qwen3-VL-8B-Instruct [2], a state-of-the-art open-source VLM capable of multi-image input. The model is fine-tuned using LoRA [8] applied to all attention projection layers ($\mathbf{q}, \mathbf{k}, \mathbf{v}, \mathbf{o}$) and MLP projection layers (`gate`, `up`, `down`), with rank $r = 16$, scaling factor $\alpha = 32$, and dropout rate 0.05. The model is trained on the geometric reasoning chain dataset using standard next-token prediction loss. Special formatting tokens demarcate the chain-of-thought segment from the final numerical output.

Training is performed for 1 epoch over 62,834 RealEstate10K sequences (from 30,702 source videos), using a cosine learning rate of 1×10^{-4} with 10% warmup and an effective batch size of 24 across 3 GPUs. Training completes in 7,062 steps (72.8 hours), reaching a final eval loss of 0.2238.

3.4. Reinforcement Learning with Geometry Rewards

To further improve numerical accuracy beyond SFT, we apply GRPO with geometry-grounded reward signals, building on the SFT-initialized model.

The reward function is $r = r_{\text{fmt}} + r_{\text{val}} + r_{\Delta R} + r_{\Delta t}$, where r_{fmt} (weight 0.2) rewards well-formed output struc-

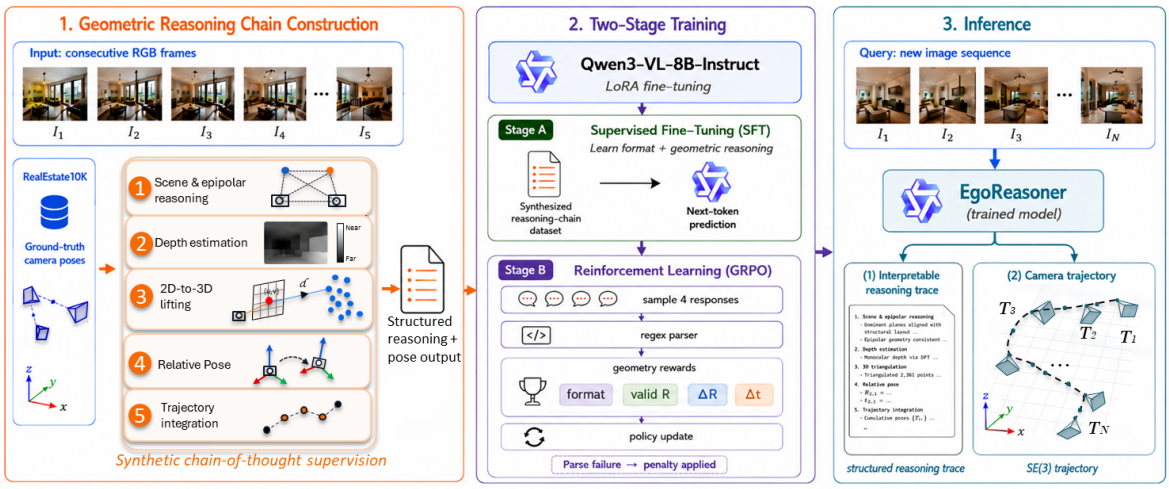

Figure 1. **EgoReasoner pipeline overview.** Given a sequence of consecutive RGB frames, the model produces a five-step geometric chain-of-thought—covering epipolar reasoning, depth estimation, 3D triangulation, relative pose estimation, and trajectory integration—before outputting the final numerical pose trajectory. The base VLM is first fine-tuned via SFT on synthesized geometric reasoning chains, then further refined by GRPO reinforcement learning with geometry-grounded reward signals.

ture, r_{val} (weight 0.1) rewards a valid rotation matrix, $r_{\Delta R}$ (weight 0.35) rewards rotation accuracy, and $r_{\Delta t}$ (weight 0.35) rewards translation accuracy. A strong parse failure penalty of -1.0 is applied when the output cannot be parsed, ensuring that the RL stage does not degrade the structured reasoning format established by SFT.

The RL training loop proceeds as follows: given 5 consecutive frames and intrinsic matrix $\mathbf{K}$, the model generates $n{=}4$ candidate responses per prompt; a regex-based parser extracts $\mathbf{R}$ and $\mathbf{t}$; the reward function scores each candidate; and GRPO updates the policy from relative reward differences. Training uses a learning rate of 1×10^{-6}, KL coefficient 0.01, temperature 0.7, and runs on 492 sequences for 1 epoch.

4. Experimental Evaluation

4.1. Dataset

EgoReasoner is trained on two non-overlapping splits of the **RealEstate10K** dataset, a large-scale collection of indoor video sequences with ground-truth camera poses. The **SFT split** comprises 62,834 five-frame sequences from 30,702 source videos; the SFT model is trained on the training portion of this split, and 60 held-out sequences are reserved for step-by-step chain-of-thought evaluation. The **RL split** comprises 74,027 usable five-frame sequences from 13,532 source videos, with *zero* video overlap with the SFT split; 492 sequences are used for RL training, and a **200-sequence** holdout is reserved for the main comparative evaluation on which all four methods are assessed. Together the two splits cover 684,455 frames in total.

Metric	Base Qwen3	Gemini 3.1 Pro†	Ours (SFT)	Ours (SFT+RL)
Parse Rate (%)	31.0	59.0	**100.0**	**100.0**
ATE median*	N/A	0.0113	0.0086	**0.0082**
RRE median (°) ↓	4.87	2.72	1.35	**1.21**
RRE mean (°) ↓	6.32	2.91	1.79	**1.71**
RTE median ↓	0.2607	0.1895	0.0308	**0.0283**
TDE median (°) ↓	80.33	18.76	16.45	**15.47**
TDE mean (°) ↓	69.63	37.33	25.25	**23.69**
TDE ≤15° (%) ↑	8.1	43.2	45.9	**48.6**

Table 1. **Camera pose estimation on the 200-seq RealEstate10K holdout.** Bold: best per metric. †Gemini prompted with geometric hints. *Base Qwen3 ATE unavailable: model outputs only 1 image-pair pose (2 frames) vs. the required 5-frame trajectory, causing frame-count mismatch.

4.2. Baselines and Metrics

We compare against two baselines: **Base Qwen3-VL-8B-Instruct** [2] (the unmodified base VLM, no geometric fine-tuning) and **Gemini 3.1 Pro** [5] prompted with explicit geometric reasoning hints. Pose accuracy is evaluated with four metrics: Relative Rotation Error (**RRE**, degrees), Relative Translation Error (**RTE**, normalized magnitude), Absolute Trajectory Error (**ATE**, RMSE after rigid alignment), and Translation Direction Error (**TDE**, degrees). For CoT evaluation we additionally report **parse rate**, epipolar error, depth relative error, and 3D triangulation error at each reasoning step.

4.3. Comparative Evaluation

Three findings stand out from Table 1, where pose metrics are computed on the commonly parsed subset. First, the

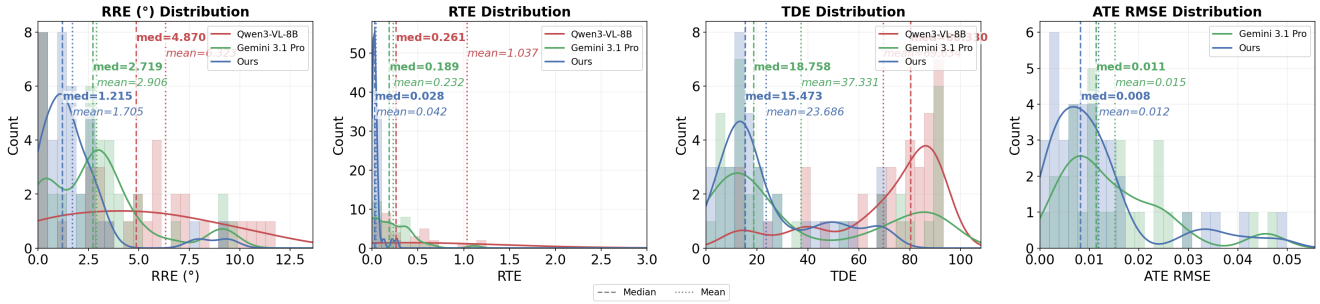

Figure 2. **Error distributions on the commonly parsed evaluation subset (200-sequence holdout).** EgoReasoner (SFT+RL) achieves consistently tighter distributions across all four metrics. Base Qwen3 ATE is absent due to frame-count mismatch (Table 1).

Step	Metric	Base Qwen3	Gemini 3.1 Pro	Ours
S1 Epipolar	Parse (%)	85.0	83.3	68.3
	Epipolar (px) ↓	6.34	2.30	**2.22**
S2 Depth	Parse (%)	76.7	5.0	**96.7**
	Depth Err ↓	3.261	1.779	**0.760**
S3 3D Tri.	Parse (%)	70.0	1.7	**81.7**
	3D Err ↓	7.430	59.436	**1.044**
S4 Rel. Pose	Parse (%)	51.7	0.0	**86.7**
	RTE mean ↓	0.7616	N/A	**0.0741**
S5 Final Traj.	Parse (%)	76.7	3.3	**60.0**
	RRE med. (°) ↓	3.32	3.99	**2.21**
	RTE med. ↓	0.1013	0.1344	**0.0596**

Table 2. **Step-by-step CoT evaluation** on the 60-seq SFT holdout. Each step is evaluated independently. Bold: best per metric.

base model's TDE median of 80.33° reveals near-random translation direction estimation, confirming that metric geometric reasoning does not emerge naturally in VLMs without geometric supervision—despite a reasonable parse rate. Second, Gemini 3.1 Pro, despite being a much larger model prompted with explicit geometric hints, parses only 59% of sequences and still yields poor translation estimation (RTE median 0.190), demonstrating that prompt engineering alone cannot substitute for geometric training. Third, EgoReasoner (SFT) achieves full parse reliability and surpasses Gemini by 6.1× in RTE, confirming that geometric chain-of-thought supervision instills both structured output behavior and substantially improved metric reasoning; error distributions are shown in Figure 2.

4.4. Chain-of-Thought Step-by-Step Evaluation

To assess whether EgoReasoner learns genuine intermediate geometric reasoning rather than learning to produce superficially formatted outputs, we evaluate each step of the reasoning chain independently on the 60-sequence SFT evaluation set. Table 2 reports parse rates and accuracy metrics at each of the five reasoning steps.

The step-by-step results reveal a key structural finding: Gemini's parse rate collapses to near zero at Steps 2–4, indicating its occasional successful final outputs in Table 1 are heuristic guesses rather than the product of structured geometric reasoning. In contrast, EgoReasoner maintains high parse rates across all intermediate steps and achieves substantial accuracy gains at each stage. The largest improvements occur at depth estimation (4.3× over the base model) and 3D triangulation (7.1×), which are the geometric foundations that the final trajectory estimate depends on—explaining why intermediate CoT supervision translates directly into better end-to-end pose accuracy.

4.5. Effect of Reinforcement Learning

The GRPO RL stage yields consistent gains across all metrics (Table 1), with translation-related metrics improving more than rotation—RTE improves by 8.1% versus 10.4% for RRE—consistent with the reward function's equal weighting of rotation and translation accuracy and RL's capacity to exploit reward gradients that SFT supervision alone cannot provide. The distribution plots in Figure 2 corroborate this: the SFT+RL curves shift most noticeably in RTE and TDE tails, indicating RL corrects the harder-to-learn translation direction errors. Importantly, parse reliability is fully preserved across all evaluation sequences, confirming that the parse failure penalty effectively prevents the RL stage from sacrificing output format for score.

5. Conclusion

We presented **EgoReasoner**, an SFT+RL framework that trains VLMs to estimate camera pose trajectories via a five-step geometric chain-of-thought reasoning process. On a 200-sequence holdout set, EgoReasoner achieves a median RRE of 1.21°, a median RTE of 0.028, and TDE $\leq$15° for 48.6% of samples—surpassing Gemini 3.1 Pro in translation accuracy by over 6× and achieving 3.2× higher parse reliability than an unmodified 8B VLM—demonstrating that metric geometric reasoning can be effectively instilled in VLMs through structured training. This capability is foundational for embodied agents in long-horizon interaction tasks, where a single VLM backbone must jointly handle semantic scene understanding and precise metric spatial self-localization.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 2, 3

[3] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13537–13547, 2024. 2

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2024. 2

[5] Google DeepMind. Gemini 3.1 pro, 2025. Accessed: 2026-05-08. 1, 3

[6] 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

[7] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems*, 2023. 2

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

[9] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 2

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 1, 2

[11] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[12] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 1

[13] 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, 2

[14] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 1, 2

[15] 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. 2

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2

[17] Lihe Yang et al. Depth anything v3. *arXiv preprint*, 2025. 2