Where Am I? Semantic Map Grounding via Vision-Language Models for Multi-Modal Localization

Suraj Borate[1], Aarav Shah[2] and Madhu Vadali[3]

***Abstract*— We address robot localization in GPS-denied indoor environments by reframing it as a *semantic reasoning* task rather than a geometric estimation problem. Motivated by the way humans localize using object-level cues and a labeled map, we ask: can a vision-language model (VLM), given a front camera image, a polar LiDAR scan, and a top-down semantic grid map, infer the robot's pose?**

We fine-tune Qwen2.5-VL-7B with LoRA and attach a lightweight regression head that predicts continuous pose coordinates (x, y, θ) directly from the model's final hidden state, bypassing text generation entirely. Training uses a composite position-and-direction loss with curriculum learning on a custom Gazebo simulation dataset (120,112 samples, 527 scenes).

On the in-distribution test set (18,017 samples), the model achieves 98.23% position accuracy (PA), 98.00% direction accuracy (DA), 96.75% full pose accuracy (FPA), a mean position error of 0.11 m, and a mean orientation error of 5.7° at 0.62 s per sample. Position accuracy drops by only 7.2% (absolute) on 7 unseen object categories (90.99%), supporting genuine semantic spatial reasoning rather than appearance memorization. When maps are incomplete, fine-tuning recovers performance to 93.72% PA—demonstrating adaptability to stale or partial map information.

Two ablations highlight cross-modal complementarity. When LiDAR is removed entirely (camera and map only, Exp. 9), PA remains at 95.06%, just 3.2% below the full system. However, when the camera provides no visible objects (wall-facing view, Exp. 6), LiDAR sustains PA at 92.33%—compared to 70.74% with neither LiDAR nor visible objects (Exp. 7)—demonstrating that LiDAR is the primary localization signal precisely when camera semantics are unavailable and acts as a reliable fallback under occlusion or sparse layouts.

I. Introduction and Problem Formulation

Classical robot localization-SLAM [2], particle filters [1], and metric visual localization [3], [4]-builds pose estimates by accumulating geometric constraints over time. These methods perform well with dense, consistent maps but degrade under perceptual noise, map incompleteness, or the absence of GPS.

Humans localize differently. Entering an unfamiliar room, a person identifies distinctive objects, estimates their distances, consults a labeled floor plan, and triangulates a position; if one cue fails, others compensate. This *semantic reasoning* over object-level priors is precisely what large VLMs are trained to support. Recent work shows that VLMs such as GPT-4V [7], LLaVA [8], and Qwen-VL [15] can describe spatial relationships, reason about viewpoints, and cross-reference visual and textual information. Transformer attention over long token sequences supports implicit multi-step inference [11], [12] the same deliberative process underlying human map reading.

We ask: *can a VLM, given a labeled semantic map, a camera image, and a LiDAR scan, localize a robot the way a human would?* Our hypothesis is that fine-tuning unlocks the VLM's latent capacity to cross-reference visual observations against map structure and output a continuous pose estimate. Related semantic localization methods use object-level maps [5] or scene graphs [6] but rely on hand-crafted geometric detectors rather than learned reasoning. VLMs have been applied to navigation [9] and robot control [10], but their use for *continuous pose regression* from multi-modal data of polar LiDAR image, front camera image and semantic map inputs is novel.

Problem statement. A robot operates in a known indoor environment represented by a 2D semantic grid map M, where each occupied cell is labeled with an object category (e.g., *toolbox*, *shoe*). At inference time the robot observes: (i) $I \in \mathbb{R}^{H \times W \times 3}$, a front-facing RGB image; (ii) S, a polar LiDAR scan rendered as a 2D bird's-eye-view image with metric scale; and (iii) M. The goal is to predict $\hat{p} = (\hat{x}, \hat{y}, \hat{\theta})$-grid position and heading $\theta \in [0, 360)$-in a **single shot**, without access to prior pose or motion history.

Contributions. We present:

- A formulation of indoor localization as single-shot multi-modal semantic reasoning, using a labeled grid map as a symbolic prior analogous to a floor plan.
- A modified Qwen2.5-VL-7B architecture with LoRA fine-tuning and a PoseHead regression module that replaces text generation with direct (x, y, θ) prediction.
- A custom Gazebo dataset (120,112 samples, 527 scenes) with five controlled evaluation splits.
- Systematic ablations across different experimental conditions, including unseen objects during training, partial maps, no objects in camera view(reliance on LiDAR and map), camera+map-only (no LiDAR polar image) and different environments.

II. Method

A. Architecture

Fig. 1 shows our pipeline. As shown in 2 three images-semantic map, front camera, and LiDAR scan-are concatenated into a multi-image prompt for Qwen2.5-VL-7B-Instruct [15]. We bypass text generation by suppressing the

*This work is supported by IIT Gandhinagar and Prime Ministers Research Fellowship

[1]Suraj Borate, PhD Student, IIT Gandhinagar, Gandhinagar, Gujarat, India surajb@iitgn.ac.in

[2]Aarav Shah, Undergraduate Student, IIT Gandhinagar, Gujarat, India aarav.shah@iitgn.ac.in

[3]Madhu Vadali, Associate Professor, IIT Gandhinagar, Gujarat, India madhu.vadali@iitgn.ac.in

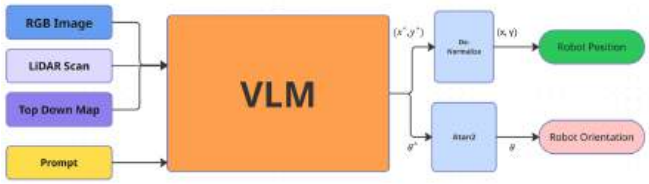


Fig. 1: System pipeline. Three input modalities are concatenated as a multi-image prompt to Qwen2.5-VL-7B fine-tuned with LoRA. The last non-padding hidden state feeds a 4-output PoseHead that produces (x, y) and direction cosines $(\cos\theta, \sin\theta)$.

full-vocabulary logit tensor (`logits_to_keep=1`) and instead attach a **PoseHead** to the pooled last non-padding hidden state $h \in \mathbb{R}^{3584}$. The head applies LayerNorm, a GELU-activated linear projection to 512 dimensions, dropout (0.1), and a final linear layer to 4 outputs:

$$\begin{aligned} h' &= \text{Linear}_{d\to 512}(\text{GELU}(\text{LayerNorm}(h)))\,, \\ \hat{p} &= \text{Linear}_{512\to 4}(h'), \quad \text{dropout} = 0.1. \end{aligned} \tag{1}$$

The output is split into normalized position $(\hat{x}_n, \hat{y}_n)$ and direction cosines $(\hat{d}_x, \hat{d}_y)$.

B. Pose Representation

Grid positions $(x, y) \in [2, 7]$ are normalized as $x_n = (x - 4.5)/2.5$. Orientation is encoded as a unit vector $d = (\cos\theta, \sin\theta)$ to eliminate the 0/360 discontinuity; heading is recovered at inference via $\hat{\theta} = \text{atan2}(\hat{d}_y, \hat{d}_x)$.

C. Training

LoRA fine-tuning. We apply LoRA [14] (r=32, α=32) to all four module groups (vision encoder and language decoder, attention and MLP layers), keeping base weights frozen. Training uses AdamW (lr $= 2\times10^{-5}$, weight decay 0.01), effective batch size 16, BF16 precision, on a single NVIDIA RTX 5090 (32 VRAM GB), via the Unsloth library.

Loss function. The composite loss is:

$$\mathcal{L} = \alpha\,\mathcal{L}_{\text{pos}} + \beta\,\mathcal{L}_{\text{dir}} + \lambda\,\|(\hat{x}_n, \hat{y}_n)\|_2^2,$$

where $\mathcal{L}_{\text{pos}} = \|(x_n^*, y_n^*) - (\hat{x}_n, \hat{y}_n)\|_2^2$ (MSE), $\mathcal{L}_{\text{dir}} = 1 - d^* \cdot \hat{d}$ (cosine distance), and the L_2 regularization term discourages predictions from drifting to the map boundary under ambiguous inputs. Weights are set to α=12.0, β=1.0, λ=0.001; the large α reflects the primacy of position accuracy in localization.

Curriculum learning. Training proceeds in three stages of increasing difficulty. In the first stage, the model is fine-tuned on the full three-modality input (front camera, LiDAR scan, and complete semantic map), establishing a strong baseline for discrete pose prediction. In the second stage, training continues on incomplete maps containing unseen object categories, encouraging the model to reason from partial evidence rather than relying on complete map-observation correspondence. In the third stage, LiDAR polar images are withheld and the model is trained on camera and map alone on data having atleast one object in view; this ablation reveals that when at least one object is visible in the camera frame, the model retains competitive localization accuracy, confirming that semantic visual cues are themselves sufficient for coarse pose estimation.

D. Evaluation Metrics

Physical coordinates are recovered as $\hat{x}_m = (s\hat{x}_n + c) \times g$, grid size g=0.57 m/cell. We report: **PA** (Position accuracy (PA) is defined as the fraction of samples for which the predicted grid cell exactly matches the ground-truth cell. Each prediction is scored as a binary outcome: 1 if the predicted (x, y) falls within the correct grid cell, and 0 otherwise. PA is then the mean of these binary scores across all evaluation samples), **DA** (direction accuracy, correct upto 45 bin), **FPA** (both correct), **MPE** (mean position error in metres), and **MOE** (mean orientation error in degrees).

III. Dataset

We construct a custom dataset using the Gazebo simulation environment. The environment is a 6×6 indoor arena subdivided into an 8×8 semantic grid (border cells = walls), leaving a 6×6 navigable interior. For each of 527 scene configurations, 8-10 objects are drawn from a 10-category training pool (shoe, bag, cone, toolbox, coke can, ball, chair, drawer, backpack, fire hose) and placed at random interior cells. Views are captured from the center of every navigable cell at 8 discrete headings (N, NE, E, SE, S, SW, W, NW), yielding 288 samples per scene and **120,112 training samples** in total-forming a fully discrete pose dataset. At each viewpoint we capture a front RGB image, a 2D polar LiDAR scan (rendered with metric scale), and the corresponding labeled grid map.

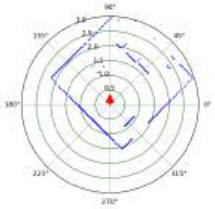
Evaluation splits. Five evaluation conditions are constructed: *In-distribution* (ID, 18,017 samples, full map, training object categories); *Unseen objects* (UO, 9,816 samples, 7 unseen object categories annotated on the map); *Partial map* (PM, 1,498 samples, 50% of objects absent from the map, simulating a stale map); *Partial map fine-tuned* (PM-FT, 4,792 samples, model retrained with PM augmentation);*Wall Facing* (991 samples, camera scenes in which only wall appear in view, forcing reliance on LiDAR and map), *No object visibility* (NV, 745 samples, camera scenes in which no objects appear in view, forcing reliance on LiDAR and map). *Camera only* (Don't provide LIDAR images, camera images with atleast one object in view. finetuned ID model on 60,601 samples and then tested it on 10,679 samples). *Camera only Fine Tune* (Don't provide LIDAR images, camera images with atleast one object in view. finetuned base model on 60,601 samples and then tested it on 10,679 samples).

IV. Experiments and Results

We conduct six completed experiments and two ongoing ones. **Exp. 1** tests Qwen2.5-VL-7B-Instruct (base model) under zero-shot prompting. **Exp. 2** fine-tunes the base model on full ID training split and evaluates on the ID test split. **Exp. 3** evaluates the Exp. 2 checkpoint on the UO split (no

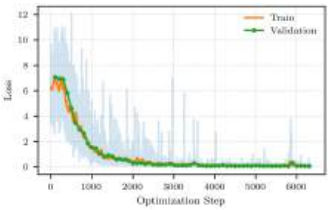
(a) Semantic map

(b) Front camera

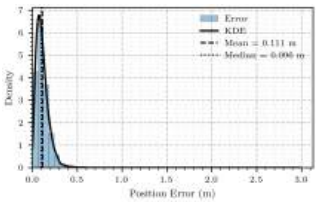


(c) LiDAR scan

Fig. 2: Example multi-modal input (scene 1357, cell C4, heading SE). Blue cells on the map denote labeled objects. LiDAR distances are annotated in metres.



(a) Train/val loss (Exp. 2)



(b) CDF of position error

Fig. 3: (a) Loss converges cleanly; validation closely tracks training, indicating no overfitting. (b) Over 90% of in-distribution predictions fall within 0.57 m (one grid cell).

retraining). **Exp. 4** evaluates on the PM split. **Exp. 5** retrains with PM augmentation. **Exp. 6** evaluates Exp2 checkpoint on scenes with robot just next to the wall facing it. **Exp. 7** evaluates Exp2 checkpoint on no objects in the camera's view. **Exp. 8** fine-tunes Exp2 checkpoint on camera and map data only (no LiDAR) using curriculum learning philosophy. **Exp. 9** evalutes the base model fine tuned on camera images and lidar data from scratch. **Exps. 10** (Environment shift) evaluates Exp. 2 checkpoint on samples across 4 different environments generated using Google's NanoBanana. are ongoing as shown in 4. All experiments use Qwen2.5-VL-7B-Instruct [15] with LoRA (r=32, α=32) on a single NVIDIA RTX 5090.

Results are summarized in Table I and Fig. 3.

Key observations.

Fine-tuning is essential (Exp. 1 vs. 2). Zero-shot prompting—including chain-of-thought variants—yields near-chance localization accuracy: the untuned VLM parses the task format (e.g., `Grid cell: D4, Orientation: E`) correctly but lacks the spatial grounding to solve it. After fine-tuning, in-distribution performance reaches PA = 98.23%, FPA = 96.75%, MPE = 0.11 m, MOE = 5.7°, at 0.62 s per sample.

TABLE I: Localization results. PA = Position Accuracy, DA = Direction Accuracy, FPA = Full Pose Accuracy, MPE = Mean Position Error (m), MOE = Mean Orientation Error (°).

Condition	**PA** (%)	**DA** (%)	**FPA** (%)	**MPE** (m)	**MOE** (°)
Zero-shot (Exp. 1)	4	18	0	NA	NA
In-dist. (Exp. 2)	98.23	98.00	96.75	0.11	5.70
Unseen obj. (Exp. 3)	90.99	96.56	89.61	0.16	7.08
Partial map (Exp. 4)	72.81	84.68	70.58	0.32	14.61
Partial map FT (Exp. 5)	93.72	97.26	92.32	0.13	6.33
Wall Facing (Exp. 6)	92.33	93.81	91.23	0.15	12.06
No obj. in view (Exp. 7)	70.74	90.07	65.36	0.27	11.47
Camera + map Cur (Exp. 8)	98.84	98.29	97.41	0.104	5.16
Camera + map FT (Exp. 9)	95.06	96.70	92.64	0.15	6.91
Domain Shift (Ex. 10)	25	87.5	25	0.42	10.69

The model reasons spatially, not by memorization (Exp. 3). On 7 unseen object categories, PA falls by only 7.2% (absolute) to 90.99% and FPA to 89.61%. A model that had simply memorized object appearances would fail entirely on novel categories. The maintained accuracy indicates instead that the model learned the *spatial configuration* of map entries relative to the robot's viewpoint—mirroring the relational strategy a human would use.

Map completeness matters, but degradation is recoverable (Exps. 4–5). With 50% of map entries absent (Exp. 4), PA drops to 72.81% and MOE rises to 14.6°. Fine-tuning on partial-map data (Exp. 5) largely restores performance (PA = 93.72%, MOE = 6.3%), showing that the model can adapt to localizing from an incomplete or stale map without relearning from scratch.

LiDAR compensates when no objects are visible (Exp. 6). Even when the camera frame is object-free (wall-facing view), the model achieves PA = 92.33% and FPA = 91.23% by drawing on LiDAR geometry and map context. This underscores the importance of metric LiDAR information precisely in scenarios where camera semantics are uninformative.

Without LiDAR, object-free views cause the sharpest accuracy drop (Exp. 7). Removing visible objects while also withholding LiDAR (camera and map only, no objects in view) reduces PA to 70.74% and FPA to 65.36%. Comparing this against Exp. 6 (92.33% PA with LiDAR under the same visibility condition) quantifies LiDAR's contribution in the hardest regime: a 21.6% absolute recovery may be attributable solely to the addition of metric range information.

Adding a current-map snapshot yields the best overall performance (Exp. 8). When the model receives a current-state map alongside the camera and LiDAR inputs, it achieves the highest scores across all conditions: PA = 98.84%,

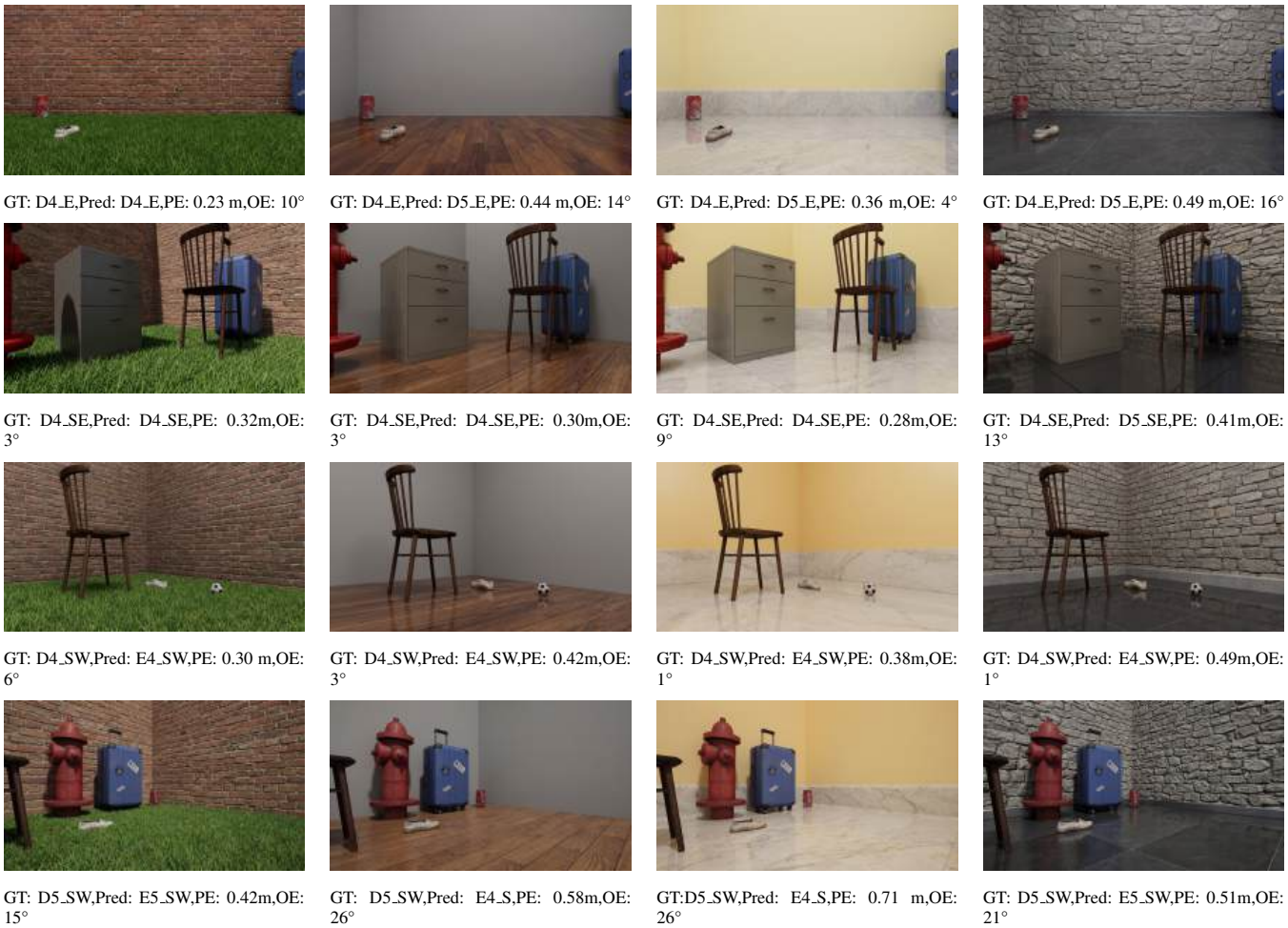


Fig. 4: Testing across environment change. Scene appearance variation is achieved using Google NanoBanana, altering lighting, texture, and color across conditions while holding geometry fixed. Each sub-image shows GT (ground-truth pose), Pred (predicted pose), PE (position error, m), and OE (orientation error, °). PA and DA defined upto 1 bin size(1 grid cell and 45 degrees bin.) is 25 % and 87.5 % respectively but PA and DA upto two bins is 100 %.

DA = 98.29%, FPA = 97.41%, MPE = 0.104 m, and MOE = 5.16°. The marginal gain over Exp. 2 (full system, in-distribution) suggests that the current map primarily reduces residual ambiguity in cluttered or symmetric scenes.

Camera and map alone are nearly sufficient under normal conditions, but not under occlusion (Exp. 9). Removing LiDAR entirely while fine-tuning on the remaining modalities (Exp. 9) yields PA = 95.06%—only 3.2% below the full three-modality system. This might seem to diminish LiDAR's role; however, the comparison with Exp. 7 (70.74% PA, no LiDAR, no visible objects) makes clear that LiDAR's contribution is conditional: when camera semantics are available, its marginal gain is modest, but when objects are absent from view it becomes the primary localization signal.

Domain shift disrupts fine-grained position but not heading (Exp. 10). Under appearance variation (textures, lighting, floor materials), PA falls to 25% while DA remains at 87.5%. Relaxing tolerances to two grid cells and two direction bins recovers both metrics to 100%, with MPE = 0.42 m and MOE = 10.69°, suggesting that appearance shift disrupts cell-level discrimination rather than coarse semantic reasoning. Although evaluated on only 16 images, this is a promising direction for targeted domain adaptation.

V. Discussion and Conclusion

We have shown that a fine-tuned VLM can localize a robot in a GPS-denied indoor environment by matching multi-modal observations to a semantic grid map, achieving 0.11 m mean position error and 5.7° mean orientation error on in-distribution data at 0.62 s per sample.

The results support four conclusions. **First**, the VLM transfers its deliberative reasoning capacity—ordinarily applied to language generation—to structured coordinate prediction via LoRA fine-tuning and a lightweight regression head, consistent with the finding that transformer attention supports implicit multi-step inference [11], [12]. **Second**, the small generalization gap to unseen categories (7.2% absolute drop in PA) confirms that the model learns *relational* spatial reasoning—identifying an object, locating it on the map, cross-referencing LiDAR distance, and triangulating—rather

than memorizing per-object appearances. This symbolic and semantic grounding is what classical geometric localization methods lack. **Third**, the no-LiDAR ablation (Exp. 9, camera and map only, 95.06% PA) and the wall-facing result (Exp. 6, 92.33% PA) together reveal the true complementarity of each modality: camera semantics drive accuracy under normal conditions, while LiDAR geometry sustains it when visible objects are absent—as further confirmed by the drop to 70.74% PA when both LiDAR and visible objects are simultaneously unavailable (Exp. 7). LiDAR is thus not a redundant modality but a targeted fallback under occlusion and sparse layouts. **Fourth**, Zero-shot domain shift experiments (Exp. 10) show that appearance variation disrupts cell-level position discrimination upto 0.42 m while leaving coarse heading estimation largely intact (DA = 87.5%), suggesting that targeted domain adaptation strategies—such as appearance-augmented fine-tuning—are a promising avenue for deployment in real environments. Approach shown in paper shows that VLMs can mimick deliberative thinking approach of human referring to map and view to figure out their current location.

Limitations. The system assumes a static, pre-built semantic map; dynamic objects and the sim-to-real transfer gap remain open challenges. Although we have tested generalization on small number of environment images, the results are promising as the accuracy is good upto two bins suggesting larger scale training required for sim2real transfer. Also we haven't trained across different camera resultions and this is part of future work.

Future work includes sim-to-real transfer on a physical robot, online map updating, richer map representations (scene graphs, 3D semantic maps), and these results motivates us for extension of this approach to continuous pose spaces without grid discretization.

Resources. The code and the resources are avialalable at where-am-i.

ACKNOWLEDGMENT

Authors thank Prime-Minister's Research Fellowship for supporting first author's research and IITGN Robotics Labs for providing resources. We also thank IIT Gandhinagar for unwavering support and funding. We thank Mr. Bhavish Rai B. in help for data collection. We thank Mr. Kishan Ved for discussions.

REFERENCES

[1] F. Dellaert, D. Fox, W. Burgard, and S. Thrun, "Monte Carlo Localization for Mobile Robots," *Proc. IEEE ICRA*, 1999.
[2] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. MIT Press, 2005.
[3] R. Arandjelović et al., "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," *Proc. IEEE CVPR*, 2016.
[4] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From Coarse to Fine: Robust Hierarchical Localization at Large Scale," *Proc. IEEE CVPR*, 2019.
[5] N. Atanasov, M. Zhu, K. Daniilidis, and G. J. Pappas, "Localization from Semantic Observations via the Matrix Permanent," *Int. J. Robotics Research*, vol. 35, no. 1-3, 2016.
[6] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization," *Proc. RSS*, 2022.
[7] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, 2023.
[8] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *Proc. NeurIPS*, 2024.
[9] D. Shah, B. Osiński, B. Ichter, and S. Levine, "LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action," *Proc. CoRL*, 2023.
[10] A. Brohan et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," *Proc. CoRL*, 2023.
[11] A. Vaswani et al., "Attention Is All You Need," *Proc. NeurIPS*, 2017.
[12] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Proc. NeurIPS*, 2022.
[13] T. Brown et al., "Language Models are Few-Shot Learners," *Proc. NeurIPS*, 2020.
[14] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," *Proc. ICLR*, 2022.
[15] Qwen Team, "Qwen2.5-VL Technical Report," Alibaba Group, 2024.