LangLoc: "Tell Me What You See"

Shaurya Kishore Panwar[1,2⋆], Roham Zendehdel Nobari[1,2*], Shirley Feng Yi Lau[1,2*], Abu Bakr Rahman Shaik[1,2*], Manuel Günther[2], Marc Pollefeys[1,3], and Daniel Barath[1,4]

[1] ETH Zürich, Switzerland
[2] University of Zürich, Switzerland
[3] Microsoft
[4] HUN-REN SZTAKI, Hungary

Abstract. We tackle fine-grained indoor localization from natural language: given a free-form description of one's surroundings, estimate the observer's 2D position and heading within a known 3D environment. Language queries are lightweight, privacy-preserving, and need no camera – yet prior work stops at coarse scene retrieval and cannot resolve an intra-scene pose. We close this gap with *LangLoc*, a three-stage pipeline that (i) retrieves the correct scene via a dual-branch GATv2 encoder with CLIP semantic features, surpassing the previous best by 8 percentage points in Top-1 recall; (ii) estimates position and heading by scoring a dense floor grid through ray-cast object visibility, reaching a median error of 0.95 m; and (iii) resolves residual ambiguity through a Bayesian dialog module that asks targeted yes/no questions and updates a pose posterior until the location is pinpointed. To support this task we contribute a benchmark of 13,000+ pose-indexed natural-language descriptions over 1,300+ indoor 3D scans. Code and data will be released. Project page: `https://rzninvo.github.io/Lang-Loc/`.

1 Introduction

Knowing where you are is fundamental to almost every location-aware service: indoor navigation, robot assistance, augmented reality, and emergency response all require an accurate pose estimate.

The dominant localization paradigm today is visual: a device captures an image or video stream, uploads it to a server, and receives a pose estimate in return [32, 36]. While effective, this approach carries significant drawbacks. Image transmission is bandwidth-heavy, especially indoors where frequent queries are needed. More critically, it is *privacy-invasive*: photos of homes, offices, and hospitals inevitably capture sensitive information that users may not wish to share. Finally, capturing a *useful* image is itself non-trivial – a photo of a blank wall carries little discriminative information, requiring users to know how to frame an informative shot.

⋆ Equal contributions.

Language offers a compelling alternative. Telling a system "I'm standing in front of a bookshelf, with a blue sofa on my left and a TV across the room" is natural, fast, and transmits almost no personally identifiable information. A text description is orders of magnitude smaller than an image, requires no camera or special hardware, and mirrors how people naturally communicate their whereabouts to one another in everyday life. This makes language localization natural in camera-prohibited but digitally-twinned settings such as hospitals, labs, and emergency dispatch. Beyond localization, human-to-agent communication also requires grounding free-form verbal goals (e.g., "go to the bookshelf and find the red book") into precise 3D poses for robots, drones, and AR assistants.

Despite this appeal, language-based localization remains largely unsolved. Existing methods address only *coarse scene retrieval* – identifying which room in a database a description refers to [14, 23]. Resolving a precise pose *within* a scene from language is an open problem: many viewpoints within the same room share similar semantics, differing only in subtle geometric or visibility cues that are difficult to capture in plain text.

We present *LangLoc*, the first pipeline for fine-grained indoor localization from natural language. Given a free-form description and a database of 3D scenes, LangLoc first retrieves the correct scene – surpassing the prior state of the art by 8 percentage points in Top-1 recall – and then estimates a 2D floor position and heading within it, achieving approximately 1 m median position error. When a description is ambiguous, the system enters an interactive dialog: it asks targeted yes/no questions (*e.g.* "Is there a chair to the left of the table?") and updates a Bayesian pose posterior until the location is resolved.

Contributions.

- **Scene retrieval.** A dual-branch GATv2 encoder with CLIP features that sets a new SOTA, with an 8 percentage points improvement over prior work [14].
- **Fine-grained localization.** A visibility-based floor-grid scoring method that estimates a 2D position and heading direction from language, achieving $\approx$1 m median error.
- **Dialog-based disambiguation.** An interactive Bayesian refinement module that resolves ambiguous descriptions through targeted yes/no questions.

2 Related Work

Visual localization. Estimating the 6-DoF camera pose in a known environment is a long-standing problem in computer vision. Structure-based methods build an explicit 3D map and localize by establishing 2D–3D correspondences followed by PnP solving [32, 36, 37]. Learned local features [16] and matching networks [33] have substantially improved robustness. Hierarchical pipelines that first retrieve candidate images then perform local matching [32] define the dominant paradigm. An alternative line of work regresses pose directly from a single image via a CNN [12, 21], trading some accuracy for architectural simplicity. Scene coordinate regression methods [9–11] recover much of this accuracy gap by

predicting dense 3D coordinates and integrating a differentiable PnP+RANSAC solver, achieving state-of-the-art single-image localization. Visual place recognition [3, 5, 8, 18] tackles the related retrieval problem of identifying the closest location in a database, analogous to the coarse stage of our pipeline but with an image query. Recent work has moved toward lighter map representations: PixLoc [35] refines poses via learned feature-metric alignment, and OrienterNet [34] localizes against 2D public maps. All of the above methods require a visual query. Our work departs from this paradigm by replacing the image with a natural-language description, probing how much localization accuracy is attainable from language alone.

Language-based and cross-modal localization. Text2Pos [23] first studied text-to-3D localization by learning a joint embedding between free-form descriptions and large-scale outdoor point clouds for coarse cell retrieval. In the indoor domain, Chen *et al*. [14] frame the problem as scene retrieval over 3D scene graphs: a text query is parsed into a graph and matched against a database of 3DSSG [40] graphs to identify the correct room. Their Text2SGM model demonstrates that scene-graph matching from language is feasible, but it stops at coarse scene identification and does not resolve an intra-scene camera pose. SceneGraphLoc [26] performs cross-modal coarse localization by matching image features to 3D scene graphs with a dual-branch GATv2 encoder – an architecture we adapt for text-to-graph retrieval. SGAligner [31] learns multi-modal scene-graph embeddings that align 3D, image, and text modalities, further demonstrating the utility of scene graphs as a bridge between modalities. Our work builds on these foundations but extends the pipeline beyond scene retrieval to fine-grained pose estimation within the retrieved scene.

Dialog-based localization and embodied navigation. In embodied settings, agents often interact with humans via dialog to resolve spatial ambiguity. Vision-and-Dialog Navigation [38] introduces cooperative dialogs for object-goal navigation. DiaLoc [43] proposes iterative dialog-based localization, where an agent narrows its pose estimate through clarification questions. The broader VLN literature [4, 19, 29] studies instruction-following in photorealistic environments. Recent LLM-based agents [44] bring explicit reasoning to navigation. These works motivate our dialog-based disambiguation module, which selects targeted questions to reduce uncertainty over candidate poses.

3D scene graphs and vision-language grounding. Scene graphs [6] encode objects, attributes, and spatial relationships in a unified hierarchical structure. 3DSSG [40] extends this idea to learned prediction of semantic scene graphs from 3D indoor reconstructions, while incremental methods [42] enable on-the-fly graph construction from RGB-D streams. Open-vocabulary extensions using foundation models [17, 22, 28] have further broadened the applicability of scene graphs by removing the need for a fixed vocabulary. On the language side, ScanRefer [13] and ReferIt3D [1] pair referring expressions with 3D bounding boxes. ScanQA [7] and SQA3D [25] extend this to question answering with spatial reasoning. Pre-trained 3D–language models [20, 45] align point clouds or multi-view features with text at scale. These resources target grounding or retrieval rather

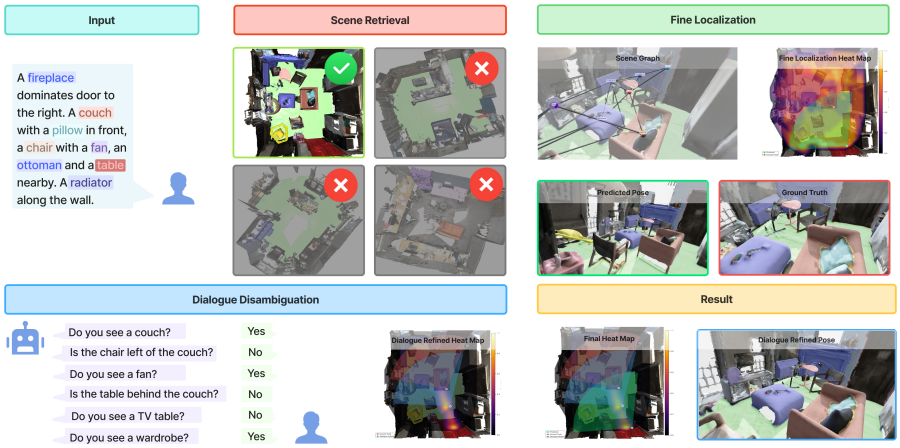

Fig. 1: LangLoc overview. *Input:* a free-form description mentioning objects such as fireplace, couch, pillow, and fan. *Scene retrieval* (Sec. 3.2 ranks candidate 3D scenes; the correct scene (green, ✓) is identified from a pool of candidates. *Fine localization* (Sec. 3.3) scores a dense floor grid by ray-cast object visibility, producing a heat map over candidate positions; the predicted pose (green frustum) is compared against the ground truth (red frustum). *Dialog disambiguation* (Sec. 3.4) asks targeted yes/no questions (e.g., "Do you see a couch?", "Is the chair left of the couch?") and iteratively refines the posterior until it concentrates on the correct pose.

than viewpoint-conditioned pose estimation, leaving a gap in both data and methods for fine-grained language-based localization that our work aims to fill.

3 Language-based Localization

Our pipeline has three stages (Fig. 1): scene retrieval (Sec. 3.2), fine localization (Sec. 3.3), and dialog-based disambiguation (Sec. 3.4).

3.1 Problem Formulation

The input is a free-form textual description T of an observer's surroundings; the output is an estimated observer pose within a known indoor environment. The setting extends language-based scene retrieval from 3D scene graphs [14]: we not only identify the correct scene but also resolve a precise intra-scene pose.

3D scene database. Let $\mathcal{D} = \{\mathcal{S}_i\}_{i=1}^{N}$ denote a database of scanned indoor scenes. Each scene $\mathcal{S}_i$ provides a reconstructed mesh with semantic instance segmentation and a 3D scene graph $G_i^s = (V_i^s, E_i^s)$. Nodes represent object instances with semantic labels; edges encode typed spatial relations (e.g., *left-of*, *on*).

Text scene graph. Following [14], we parse T into a text scene graph $G^t = (V^t, E^t)$. Nodes represent objects; edges encode relational statements from the text. This enables structured matching against the database graphs $\{G_i^s\}$.

Tasks. Given $(T, \mathcal{D})$, the system executes three tasks:

1. **Scene retrieval.** Produce a ranked list of candidate scenes $\pi(T) = (i_1, \ldots, i_K)$ by matching G^t against each G_i^s.
2. **Fine localization.** For the top candidate scene, estimate a 2D floor position $\hat{\mathbf{c}} \in \mathbb{R}^2$ and a unit direction vector $\hat{\theta} \in \mathbb{S}^2$. We denote the full pose as $\hat{\mathbf{p}} = (\hat{\mathbf{c}}, \hat{\theta}) \in \mathbb{R}^2 \times \mathbb{S}^2$ and the ground-truth pose as $\mathbf{p}^\star = (\mathbf{c}^\star, \theta^\star)$.
3. **Dialog-based disambiguation.** When the posterior over poses is multi-modal, the system asks targeted yes/no questions about object presence or spatial relations to iteratively reduce ambiguity [38, 43].

3.2 Language-based Scene Retrieval

The first stage matches the text scene graph G^t to the most compatible scene in $\mathcal{D}$ (Fig. 2). We use a dual-branch graph encoder with CLIP [30] semantic features, extending [14, 26].

Graph representation. Both 3D scene graphs and text scene graphs share the same node-edge structure, so we encode them identically. Each node v_i represents an object instance with semantic label l_i. We form a per-node feature by concatenating the object's 3D centroid $\mathbf{c}_i \in \mathbb{R}^3$, its mean RGB color $\mathbf{k}_i \in \mathbb{R}^3$, and a CLIP [30] text embedding of the label:

$$\mathbf{f}_i = [\, \mathbf{c}_i \parallel \mathbf{k}_i \parallel \phi(l_i) \,] \in \mathbb{R}^{518}, \tag{1}$$

where $\phi(\cdot) \in \mathbb{R}^{512}$ is the CLIP text encoder (ViT-B/32). For text-derived graphs, where geometry and color are unknown, we zero-pad these fields and let the network rely on semantic and relational cues alone.

We also compute a scene-level CLIP descriptor by encoding a sentence that lists all unique object labels $\mathcal{L}(G)$ in the graph:

$$\mathbf{u}(G) = \phi(\text{"A room with } l_1, \ldots, l_K\text{"}), \quad \{l_k\}_{k=1}^K = \mathcal{L}(G). \tag{2}$$

This provides a coarse, holistic summary of what objects are present, complementing the fine-grained node-level features.

Dual-branch encoder. The encoder maps any graph to an embedding $\mathbf{z}(G) \in \mathbb{R}^d$ (d=256). 3D graphs carry both geometric layout and semantic relations; text graphs carry only relations. We handle this asymmetry with two branches, merged by a learned gate [26].

Node features are first projected to the working dimension d via an MLP:

$$\mathbf{x}_i = \mathrm{MLP}(\mathbf{f}_i) \in \mathbb{R}^d. \tag{3}$$

The *geometric branch* connects each node to its k=5 nearest spatial neighbors and propagates messages conditioned on their relative position and size:

$$\begin{aligned} \mathbf{e}_{ij}^{\text{geom}} &= [\, \Delta\mathbf{c}_{ij} \parallel \|\Delta\mathbf{c}_{ij}\|_2 \parallel r_i \parallel r_j \parallel \mathbf{0}_2 \,], \\ \mathbf{g}_i &= \text{GATv2}_{\text{geom}}(\{\mathbf{x}_j, \mathbf{e}_{ij}^{\text{geom}}\}_{j\in\mathcal{N}(i)}), \end{aligned} \tag{4}$$

where $\Delta\mathbf{c}_{ij} = \mathbf{c}_j - \mathbf{c}_i$ and r_i, r_j are bounding radii. The *relation branch* embeds each relation string with CLIP. Semantically similar relations (e.g., *left_ of* and *to the left of*) are thus naturally close in embedding space:

$$\mathbf{e}_{ij}^{\text{rel}} = \phi(r_{ij}), \quad \mathbf{t}_i = \text{GATv2}_{\text{rel}}(\{\mathbf{x}_j, \mathbf{e}_{ij}^{\text{rel}}\}_{j:(i,j)\in E}). \tag{5}$$

A learned gate $\boldsymbol{\alpha}_i = \sigma(\text{MLP}([\mathbf{g}_i \| \mathbf{t}_i])) \in (0,1)^d$ fuses the two branches per node:

$$\mathbf{h}_i = \boldsymbol{\alpha}_i \odot \mathbf{g}_i + (1 - \boldsymbol{\alpha}_i) \odot \mathbf{t}_i. \tag{6}$$

For text graphs, where geometry is absent, the gate learns to rely primarily on the relation branch. Mean-pooling the node embeddings and concatenating the global descriptor yields the final graph embedding:

$$\mathbf{z}(G) = \text{MLP}\Big(\Big[\frac{1}{|V|}\sum\nolimits_i \mathbf{h}_i \;\Big\|\; \mathbf{u}(G)\Big]\Big) \in \mathbb{R}^d. \tag{7}$$

Training. We train with an InfoNCE contrastive loss [14] that pulls matching text-scene graph pairs together and pushes non-matching pairs apart. Each training pair consists of a sparse text graph G^t and the corresponding dense 3D scene graph G^s; negatives are sampled from different scenes.

Inference scoring. At test time, we rank each database scene G_i^s by a weighted combination of three complementary cues: the learned graph embedding similarity, the global CLIP descriptor similarity, and a simple label-overlap score:

$$\begin{aligned} \text{score}(G^t, G_i^s) = w_{\text{emb}} \cos(\mathbf{z}(G^t), \mathbf{z}(G_i^s)) + w_{\text{glob}} \cos(\mathbf{u}(G^t), \mathbf{u}(G_i^s)) \\ + w_{\text{jac}} \text{F1}(\mathcal{L}(G^t), \mathcal{L}(G_i^s)), \end{aligned} \tag{8}$$

where F1 measures the overlap of unique object labels and the weights w_{emb}, w_{glob}, w_{jac} are tuned on a validation split. Ranking by this score yields the retrieved list $\pi(T)$.

3.3 Fine Localization

Once scene retrieval identifies the correct environment, the second stage estimates a precise pose within it. The key insight is that the observer must be standing at a position from which the mentioned objects are *simultaneously visible and nearby*. We exploit this by matching text entities to 3D objects and then scoring a dense grid of floor positions via ray-casting against the scene mesh, counting how many matched objects are visible from each candidate.

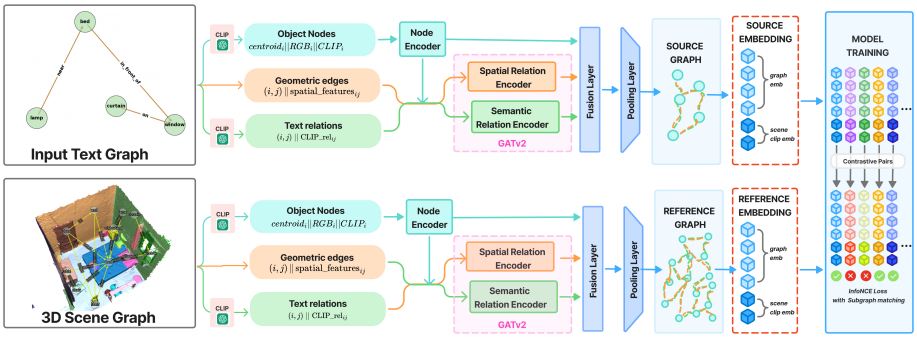

Fig. 2: Coarse retrieval pipeline overview. Given a natural language query, we first construct a scene graph representation where nodes encode object-level features (spatial attributes and CLIP embeddings) and edges capture relational constraints. This graph is processed through a dual-branch architecture that maintains separate pathways for geometric and semantic information before adaptive fusion. During training, an InfoNCE objective learns to align text-derived query graphs with 3D database graphs by learning discriminative embeddings. This simultaneously pulls together matching query-database pairs while pushing apart non-matching pairs.

Object-level matching. We embed each node label and relationship predicate into $\mathbb{R}^d$ using Word2Vec [27]. Let $\mathbf{Q} \in \mathbb{R}^{|V^t| \times d}$ and $\mathbf{S} \in \mathbb{R}^{|V^s| \times d}$ collect the ℓ_2-normalized node embeddings for query and scene nodes. We compute a cosine similarity matrix $\mathbf{M} = \hat{\mathbf{Q}}\,\hat{\mathbf{S}}^\top$. Each query node is greedily assigned to its best unused scene match above a similarity threshold.

Candidate position grid. We sample candidate 2D floor positions $\mathbf{c}_k \in \mathbb{R}^2$ on a uniform grid over the walkable floor area at spacing Δ. Each candidate represents a hypothesis for where the observer is standing. For visibility queries, we lift each candidate to 3D by adding a fixed eye height h above the floor.

Visibility scoring. For each candidate $\mathbf{c}_k$ and each matched object centroid $\boldsymbol{\mu}_o \in \mathbb{R}^3$, we cast a ray from the 3D eye position toward $\boldsymbol{\mu}_o$ on the scene mesh. The object is visible if the ray is not occluded. The score combines a visibility count with a proximity bonus that favors standing positions close to the mentioned objects:

$$s_k = \underbrace{\sum\nolimits_o \mathbb{I}[\mathrm{vis}(o, \mathbf{c}_k)]}_{\text{visibility count}} + w_d \underbrace{\sum\nolimits_{o:\,\mathrm{vis}(o,\mathbf{c}_k)} \exp\!\left(-d_{ko}/\tau_d\right)}_{\text{proximity bonus}}, \tag{9}$$

where d_{ko} is the horizontal distance from $\mathbf{c}_k$ to the object. Scores are converted to a probability distribution via softmax: $P(\mathbf{c}_k) \propto \exp(s_k/\tau)$.

Orientation estimation. For the top-scoring positions, we estimate the viewing direction $\hat{\theta}_k \in \mathbb{S}^2$ by selecting the FoV frustum orientation that captures the most visible object directions, then taking the normalized mean of those directions.

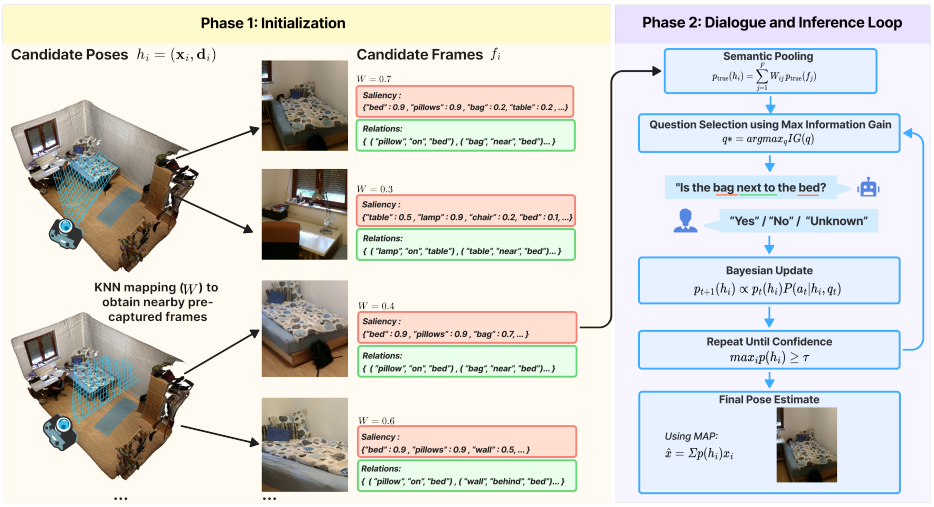

Fig. 3: Dialog-based disambiguation overview. *Phase 1 (Initialization):* candidate floor poses $h_j{=}(\mathbf{x}_j, \mathbf{d}_j)$ from the fine-localization grid are mapped via KNN to nearby pre-captured reference frames f_i, each annotated with visible object labels, spatial relations, and a matching weight w. *Phase 2 (Dialog and Inference Loop):* a semantic pooling step aggregates frame beliefs into a posterior $p(h_j)$. At each round, the system selects the yes/no question with maximum information gain (*e.g.* "Is the bag next to the bed?"), receives the user's answer, and updates the posterior via Bayes' rule. The loop repeats until the posterior concentrates above a confidence threshold τ, at which point the final pose is returned as the mode of the posterior.

Final prediction. The output pose $\hat{\mathbf{p}} = (\hat{\mathbf{c}}, \hat{\theta})$ is the weighted mean of posterior floor position paired with its estimated heading.

3.4 Dialog-based Disambiguation

When several spatially distinct viewpoints receive similar scores, a single description may not uniquely determine the pose. A room may, for instance, contain two similar seating areas, each consistent with "a sofa facing a TV". To resolve such ambiguities, we introduce a dialog module (Fig. 3) that asks targeted yes/no questions and uses the answers to iteratively narrow the set of plausible viewpoints via Bayesian posterior updates.

Frame-level posterior. We discretize the pose space into several reference frames $\{(\mathbf{t}_j, \theta_j)\}_{j=1}^{F}$, where $\mathbf{t}_j \in \mathbb{R}^2$ is the floor position and θ_j is the viewing direction. Each frame is annotated with visible labels $\mathcal{L}_j$ and spatial relations $\mathcal{R}_j$. In practice, we first *pool* a compact set of representative frames (e.g., the most frequently matched frames under a candidate-to-frame KNN mapping), and perform dialog only over this pooled set. The fine-localization belief over candidate poses is mapped to frames via a distance-weighted soft assignment matrix W,

Table 1: Scene retrieval Recall@k (%) on ScanScribe [14]. Each text query is matched against a pool of *10 candidate scenes*. All methods encode text and 3D scene graphs into a shared embedding space and rank by similarity. Best in bold.

Method	Top-1	Top-2	Top-3	Top-5
Text2Pos (from scratch)	11.63 ± 0.99	22.57 ± 1.16	33.03 ± 1.36	52.80 ± 1.67
Text2Pos (fine-tuning)	10.30 ± 0.93	20.74 ± 1.32	31.65 ± 1.45	53.42 ± 1.68
CLIP2CLIP	33.09 ± 0.21	52.53 ± 0.24	65.98 ± 0.21	82.69 ± 0.13
Text2SGM (match-prob)	63.70 ± 1.85	81.40 ± 1.85	89.31 ± 1.42	95.71 ± 0.93
Text2SGM (cos-sim)	68.27 ± 2.05	84.95 ± 1.62	91.65 ± 1.23	97.14 ± 0.76
Text2SGM (ret-based)	68.61 ± 0.04	85.00 ± 0.02	91.57 ± 0.01	96.99 ± 0.00
LangLoc (ours)	**76.70 ± 4.58**	**90.40 ± 2.73**	**96.10 ± 1.58**	**98.90 ± 1.22**

Table 2: Scene retrieval Recall@k (%) on ScanScribe [14], retrieving from the full set of *55 test scenes*. Unlike Tab. 1, the candidate pool is not subsampled, making the task more challenging. We evaluate Top 5, 10, 20 and 30 scenes. Since Tab. 1 and this table use different candidate-pool sizes (10 vs. 55) and different k ranges, Recall@k values are not directly comparable across the two tables. Best in bold.

Method	Top-5	Top-10	Top-20	Top-30
Text2Pos (from scratch)	10.27 ± 0.83	21.24 ± 1.17	39.36 ± 1.43	57.30 ± 1.48
Text2Pos (fine-tuning)	9.21 ± 0.92	18.27 ± 1.24	39.27 ± 1.61	58.63 ± 1.60
CLIP2CLIP	33.26 ± 0.21	53.47 ± 0.23	73.76 ± 0.18	86.23 ± 0.11
Text2SGM (match-prob)	70.24 ± 1.93	85.71 ± 1.38	93.36 ± 1.06	97.25 ± 0.65
Text2SGM (cos-sim)	76.34 ± 2.06	87.83 ± 1.55	95.40 ± 0.89	98.24 ± 0.57
Text2SGM (ret-based)	76.29 ± 0.16	87.77 ± 0.10	95.34 ± 0.05	98.18 ± 0.02
LangLoc (ours)	**83.30 ± 3.74**	**91.60 ± 3.01**	**97.10 ± 1.51**	**98.80 ± 0.87**

yielding the initial frame belief:

$$p_0(j) \propto (W^\top p_C)_j, \tag{10}$$

where p_C denotes the (normalized) prior over candidate poses.

Bayesian update. Each round asks a question q and receives $a \in \{\text{yes}, \text{no}, \text{unknown}\}$. We use *label* questions (object visibility) and *relation* questions (spatial configuration). For each frame j, we compute a soft truth probability $p_{\text{true}}(j) \in [0, 1]$ and answerability $p_{\text{ans}}(j) \in [0, 1]$: for labels, $p_{\text{true}}(j)$ comes from visibility/salience in frame j; for relations, $p_{\text{true}}(j)$ is determined by membership in $\mathcal{R}_j$ and $p_{\text{ans}}(j)$ by the involved objects. A reliability α flips yes/no with probability $(1 - \alpha)$, and we allocate explicit mass to `unknown` that increases as $p_{\text{ans}}(j)$ decreases. We update the posterior by

$$p(j) \leftarrow \frac{p(j)\,\mathcal{L}(a \mid j)}{\sum_{j'} p(j')\,\mathcal{L}(a \mid j')}. \tag{11}$$

Question selection. We greedily select an informative question under the current posterior. By default, we maximize expected information gain (including

Table 3: Scene retrieval Recall@k (%) on ScanScribe [14] using *LLM-generated queries* from scene images instead of scene-graph-derived text. This reduces lexical overlap between query terms and database object labels, testing generalization to natural phrasing. Best in bold.

Method	Top-1	Top-2	Top-3	Top-5
CLIP2CLIP	33.07 ± 0.21	52.52 ± 0.23	66.00 ± 0.21	82.78 ± 0.13
Text2SGM (cos-sim)	34.22 ± 1.77	56.67 ± 1.70	68.78 ± 2.35	82.11 ± 1.23
LangLoc (ours)	**59.5 ± 5.26**	**76.4 ± 4.94**	**87.8 ±3.12**	**96.2 ± 2.27**

`unknown`):

$$s(q) = H(p) - \sum\nolimits_{a\in\{\mathrm{y,n,u}\}} p(a \mid q)\, H\big(p(\cdot \mid a, q)\big), \qquad p(a \mid q) = \sum_j p(j)\, \mathcal{L}(a \mid j, q). \tag{12}$$

As a lightweight alternative, we support the balanced split heuristic

$$s_{\mathrm{bal}}(q) = -\big|p_{\mathrm{yes}}(q) - \tfrac{1}{2}\big|, \qquad p_{\mathrm{yes}}(q) = \sum\nolimits_j p(j)\, y_j, \tag{13}$$

with y_j indicating whether frame j implies a "yes". Relation questions are preferred when available; we filter out questions that are too unbalanced or insufficiently answerable under $p(\cdot)$. For label questions, we apply an IDF-based downweighting to avoid frequently occurring, low-discriminative labels.

Pose estimation. After R dialog rounds, the final pose is the mode of the posterior:

$$\hat{\mathbf{c}} = \mathbf{t}_{j^*} \in \mathbb{R}^2, \qquad \hat{\theta} = \theta_{j^*} \in \mathbb{S}^2, \qquad j^* = \arg\max_j\ p(j). \tag{14}$$

We report *Pos.* (m) as the error of $\hat{\mathbf{c}}$ in (Tab. 4).

4 LangLoc Dataset

Existing 3D vision-language datasets provide object-centric grounding [1, 13] or scene-level captions [45], but not pose-indexed egocentric descriptions suitable for localization. We build such a benchmark by extending 3RScan [39] with the pose-indexed text descriptions over 1,300+ indoor scans, with 13,000+ pose-indexed natural-language descriptions.

Keyframe selection. Raw RGB-D sequences contain many blurry, redundant, or uninformative frames. We select a compact set of high-quality keyframes per scene in three steps. First, each frame is scored by an image quality model [2] and low-quality frames are discarded. Second, we render the scene mesh to determine which objects are visible from each surviving frame and compute pairwise spatial relations (e.g., *left_of*, *above*) in camera coordinates. Third, we apply a two-stage determinantal point process (DPP) [24]: Stage 1 selects semantically

Table 4: Fine localization on 100-scene subsets of two datasets. Given the correct scene and a text description, estimate the observer's 2D floor position and heading. *Midpoint*: floor centroid. *VLM*: Qwen2.5-VL-2B [41] with top-down rendering and query (zero-shot). *Top-10 Pos.*: minimum position error among the 10 highest-scoring grid cells. *3D IoU*: frustum intersection-over-union between predicted and ground-truth views (↑). For the dialog setting, answers are obtained in a controlled protocol: by a human annotator on 3RScan and by Qwen2.5-VL-2B using ground-truth reference-frame metadata on ScanNet. Thus, the dialog rows should be interpreted as controlled dialog evaluations rather than full real-user studies. All other metrics: lower is better. Best in bold, second best underlined.

(a) 3RScan [39]

	Top-10 Pos. (m) ↓		Pos. (m) ↓		Angle (deg) ↓		3D IoU ↑
Method	**Mean**	**Med.**	**Mean**	**Med.**	**Mean**	**Med.**	**Mean**
Midpoint	–	–	<u>1.416</u>	<u>1.347</u>	–	–	–
VLM	–	–	1.582	1.420	85.48	85.54	0.062
LangLoc w/o dialog	<u>1.037</u>	<u>0.941</u>	1.712	1.551	<u>46.07</u>	<u>37.24</u>	<u>0.172</u>
LangLoc w/ dialog	**0.983**	**0.890**	**0.926**	**0.799**	**39.52**	**33.37**	**0.342**

(b) ScanNet [15]

	Top-10 Pos. (m) ↓		Pos. (m) ↓		Angle (deg) ↓		3D IoU ↑
Method	**Mean**	**Med.**	**Mean**	**Med.**	**Mean**	**Med.**	**Mean**
Midpoint	–	–	<u>1.279</u>	<u>1.098</u>	–	–	–
VLM	–	–	1.382	1.099	93.96	91.27	0.030
LangLoc w/o dialog	<u>1.254</u>	<u>1.065</u>	1.676	1.314	<u>42.67</u>	<u>34.66</u>	<u>0.236</u>
LangLoc w/ dialog	**0.532**	**0.210**	**0.593**	**0.069**	**23.77**	**5.12**	**0.593**

informative frames using a quality score that rewards object diversity and geometric complexity; Stage 2 enforces spatial diversity by penalizing frames with overlapping positions, viewing directions, and visibility masks.

Description generation. For each selected keyframe, the visible object list and spatial relations are assembled into a structured prompt for an LLM, which produces a natural-language scene description grounded at that viewpoint. These automatic descriptions are supplemented by human annotations collected via a dedicated annotation interface to form the final training and evaluation labels. Full pipeline details and hyperparameters are provided in the supp. material.

5 Experiments

We evaluate scene retrieval (Sec. 5.1) and fine localization with and without dialog (Sec. 5.2). For retrieval, we use the ScanScribe benchmark [14] built on 3RScan [39] with 3DSSG [40] annotations and follow its standard train/val/test split. For fine localization, we evaluate on the LangLoc dataset (Sec. 4), built on 3RScan, and on ScanNet [15].

Implementation details. Text descriptions are parsed into scene graphs with GPT-4o-mini. Node features concatenate centroid coordinates (3D), color attributes (3D), and CLIP ViT-B/32 embeddings (512D), totaling 518 dimensions. Scene-level CLIP embeddings are computed from object label sets and fused with graph embeddings via a two-layer MLP with LayerNorm. The dual-branch GATv2 encoder is trained for 70 epochs using AdamW (learning rate 5×10^{-4}, weight decay 10^{-4}) with a contrastive loss at temperature 0.07, 50% negative sampling, and gradient clipping (max norm 1.0). For fine localization, we sample a dense floor grid at spacing Δ=0.25 m and eye height h=1.6 m, and score each cell by single-ray visibility casting against the scene mesh.

5.1 Scene Retrieval

We report Recall@k under the three evaluation protocols of [14]. Baselines are taken from [14]: Text2Pos [23] learns a joint text–point-cloud embedding, CLIP2CLIP matches CLIP embeddings of text and rendered views, and Text2SGM performs scene-graph matching (match-prob, cos-sim, ret-based).

10-scene pool (Tab. 1). Each query is matched against 10 candidate scenes. LangLoc achieves 76.7% Top-1 recall, exceeding the strongest Text2SGM variant (68.6%) by +8.1 pp, and reaches 98.9% at Top-5. The gains across k suggest that our dual-branch CLIP-based encoder with gated fusion learns more discriminative scene embeddings than single-branch alternatives.

Full test set (Tab. 2). Retrieval over all 55 test scenes is harder (5.5× larger pool). LangLoc reaches 83.3% Top-5 and 91.6% Top-10, outperforming Text2SGM by 7 and 4 points, respectively. Text2Pos drops below 10% at Top-5, while CLIP2CLIP improves to 33.3% Top-5 but remains far below graph-based methods, highlighting the benefit of object-level structure and spatial relations.

LLM-generated queries (Tab. 3). To test robustness beyond graph-derived text, we generate queries with GPT-4o-mini from scene images and detected objects, reducing lexical overlap and increasing phrasing variability. LangLoc obtains 59.5% Top-1, a +25 pp gain over Text2SGM (34.2%) and +26 pp over CLIP2CLIP (33.1%); at Top-5 it reaches 96.2%, indicating the correct scene is almost always in the top candidates under domain shift.

5.2 Fine Localization

Given the retrieved scene, we evaluate how accurately LangLoc estimates the observer's 2D floor position $\hat{\mathbf{c}}$ and heading $\hat{\theta}$. We compare against two baselines that require no task-specific training.

Baselines. The *midpoint* baseline places the observer at the centroid of the floor bounding box; it uses no language input and does not predict a heading, so angular error and IoU are not applicable. The *VLM* baseline provides Qwen2.5-VL-2B [41] with a top-down rendering of the scene and the text query, and asks it to predict pixel coordinates and a heading in a zero-shot setting. The predicted pixel locations are unprojected to 3D via the known intrinsics and floor-plane intersection.

Metrics. We report four metrics. *Position error*: Euclidean distance $\|\hat{\mathbf{c}} - \mathbf{c}^\star\|_2$ in meters over 2D locations. *Top-k position error*: minimum distance among the k highest-scoring grid cells, providing a softer measure for multi-modal posteriors. *Angular error*: geodesic angle $\arccos\Big(\mathrm{clip}(\hat{\theta}^\top \theta^\star, -1, 1)\Big)$ between the predicted and ground-truth unit direction vectors, reported in degrees. *3D IoU*: frustum intersection-over-union between the predicted and ground-truth viewing frustums on the floor plane (higher is better).

Evaluation data. Because the dialog-based variant requires human annotations – each query involves multiple rounds of interactive disambiguation – we evaluate

Fig. 4: Qualitative frustum overlap on two datasets. We visualize ground-truth frustum (red), predicted frustum (yellow), and their overlap (orange).

on 100 randomly selected scenes from each dataset. Annotating dialog turns is time-consuming, so we reserve the full-scale evaluation (all 1,300 scans of the LangLoc dataset) for the non-dialog variant and report it separately in Tab. 6; complete per-dataset results are provided in the supplementary material.

Results on 100-scene subsets (Tab. 4). On the 3RScan split, LangLoc without dialog already halves the angular error compared to the VLM baseline (46.1° mean vs. 85.5°), demonstrating that ray-cast visibility scoring yields a meaningful heading estimate. The midpoint does not predict a heading (marked "–" in the table). Adding dialog further reduces the median position error from 1.55 m to 0.80 m – a 49% improvement – and the mean angular error from 46.1° to 39.5°. The IoU nearly doubles from 0.172 to 0.342, as the dialog module narrows the posterior and aligns the predicted frustum with the ground truth. The VLM achieves only 0.062 IoU despite predicting a heading, since its orientations are essentially random (∼85° error).

An instructive pattern emerges in the raw position error. The midpoint baseline achieves a seemingly competitive 1.42 m mean despite using *no* language at all. This is an artifact of small indoor rooms: the floor centroid is mechanically close to every point in a compact space, but it carries no practical value – it conveys neither a meaningful within-room position nor a viewing direction. Therefore, the midpoint baseline should not be interpreted as a meaningful localization result, since it provides neither orientation nor viewpoint-specific evidence, unlike LangLoc's angular, frustum-overlap, and posterior-based estimates.

LangLoc w/o dialog reports a higher raw position error (1.71 m) because it commits to a specific floor region near the described objects, which can be wrong when several object clusters match the description. This suggests that the main residual ambiguity is not heading estimation, where LangLoc is already far stronger than the VLM, but choosing among multiple spatially separated pose clusters that satisfy the same description. The dialog module targets exactly this failure mode by asking discriminative questions across clusters, thereby collapsing the posterior toward the correct viewpoint.

The Top-10 metric disambiguates the two behaviors: the correct position is almost always among the high-scoring cells, reaching 1.04 m (Top-10 mean) –

Table 5: Human dialog pilot on 10 scenes. Unlike the controlled dialog rows in Tab. 4, this pilot uses human-written descriptions and human-typed dialog answers. The *Human* row reports a person localizing from the same description. 3D IoU measures frustum intersection-over-union between predicted and ground-truth views; higher is better. All other metrics are lower-is-better.

	Top-10 Pos. (m) ↓		**Pos. (m) ↓**		**Angle (deg) ↓**		**3D IoU ↑**
Method	Mean	Med.	Mean	Med.	Mean	Med.	Mean
Midpoint	–	–	0.91	0.83	–	–	–
Qwen	–	–	1.26	0.85	78.68	75.38	0.03
LangLoc	1.16	0.96	1.74	1.19	49.93	54.08	0.21
LangLoc + Human dialog	**0.61**	**0.24**	**0.66**	**0.36**	<u>45.68</u>	<u>40.34</u>	<u>0.25</u>
Human	–	–	<u>0.70</u>	<u>0.51</u>	**20.37**	**15.31**	**0.39**

well below the midpoint's raw error. With dialog, LangLoc achieves 0.93 m mean and 0.80 m median position error, outperforming the midpoint on every metric.

On ScanNet, which features different room layouts and annotation conventions, LangLoc w/o dialog generalizes well: it achieves a Top-10 mean of 1.25 m, an angular error of 42.7°, and an IoU of 0.236 – comparable to the 3RScan split. The VLM again produces near-random headings (∼94°) and near-zero IoU (0.030). Adding the dialog on ScanNet yields even larger gains than on 3RScan: the median position error drops from 1.31 m to 0.07 m, the median angular error from 34.7° to 5.1°, and the IoU rises from 0.236 to 0.593.

Two factors compound this: ScanNet frames contain more visible objects on average (6 vs. 4), making each dialog question more discriminative, and its rooms are less spatially ambiguous, causing the posterior to collapse to a single mode in over 40% of scenes and snap the MAP estimate below grid spacing.

Here, Pos. and Top-10 Pos. reflect different estimators: the former is the mode of the posterior ($j^* = \arg\max_j p(j)$, always on a grid cell), whereas the latter is the minimum distance among the ten highest-scoring grid cells, providing a softer measure when the posterior is multi-modal. The Top-10 mean reaches 0.53 m, confirming that the Bayesian posterior concentrates sharply around the correct pose after a few clarification rounds.

Human dialog pilot. To test whether the controlled dialog gains transfer to real interaction, we additionally run a 10-scene pilot with human-written descriptions and human-typed dialog answers. LangLoc with human dialog improves over single-shot LangLoc and approaches a human asked to localize from the same description (these human baselines are in the last row of Tab. 5). While the human annotator better estimates the heading directions, LangLoc with the dialog achieves the lowest position errors.

Full-scale evaluation (Tab. 6). On the complete LangLoc dataset (1,300 scans, 13k+ descriptions), LangLoc w/o dialog achieves a Top-10 median position error of 0.95 m and a median angular error of 39.8°, consistent with the 100-scene subset and confirming that LangLoc scales to the full benchmark. Midpoint again

Table 6: Fine localization on the 1,300 scenes of the full LangLoc dataset. Given the correct scene and a text description, estimate the observer's 2D floor position and heading. *Midpoint*: floor centroid (no language). *VLM*: Qwen2.5-VL-2B [41] with top-down rendering and query (zero-shot). *Top-10 Pos.*: minimum position error among the 10 highest-scoring candidate locations. *3D IoU*: frustum intersection-over-union between predicted and ground-truth views (↑). All other metrics: lower is better. Best in bold, second best underlined.

	Top-10 Pos. (m) ↓		**Pos. (m) ↓**		**Angle (deg) ↓**		**3D IoU ↑**
Method	**Mean**	**Med.**	**Mean**	**Med.**	**Mean**	**Med.**	**Mean**
Midpoint	–	–	**1.369**	**1.259**	–	–	–
VLM (Qwen2.5-VL)	–	–	1.538	1.350	<u>84.70</u>	<u>81.70</u>	<u>0.018</u>
LangLoc w/o dialog	**1.153**	**0.951**	<u>1.534</u>	<u>1.308</u>	**46.85**	**39.80**	**0.147**

achieves a low raw position error (1.26 m median vs. 1.31 m for LangLoc) due to the small room sizes, but this remains a vacuous baseline: it provides neither a meaningful within-room position nor any viewing direction. LangLoc's 3D IoU of 0.147 is an order of magnitude higher than the VLM's 0.018, confirming that the method recovers both accurate position and orientation from language alone.

6 Conclusion

We presented LangLoc, the first pipeline for fine-grained indoor localization that estimates a 2D position and heading from natural language text description – without any image – by combining a dual-branch GATv2 retrieval encoder, visibility-based floor-grid pose scoring, and a Bayesian dialog module for disambiguation. On the proposed LangLoc dataset (13k+ descriptions over 1,300+ scans) and on ScanNet, it achieves approximately 1 m median position accuracy and meaningful headings, substantially outperforming geometry-only and zero-shot VLM baselines. Interactive dialog yields further large gains – reducing the ScanNet median error to 7 cm and 5° – showing that a few targeted yes/no questions resolve much of the spatial ambiguity inherent in a single description.

Limitations and future work. The current pipeline operates on pre-built 3D scene graphs with known object labels and relationships. Extending it to open-vocabulary or incrementally constructed graphs would broaden applicability to environments without prior annotation. Performance degrades in large, cluttered scenes where many viewpoints share similar object configurations – a challenge that richer spatial reasoning or multi-turn dialog strategies could address. Finally, while language queries are inherently privacy-preserving, the method still requires access to a detailed 3D model of the environment; relaxing this assumption, *e.g.* by localizing against coarser floor plans or schematic maps, is a promising direction for practical deployment.

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.J.: ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In: 16th European Conference on Computer Vision (ECCV) (2020)
2. Agnolucci, L., Galteri, L., Bertini, M.: Quality-aware image-text alignment for opinion-unaware image quality assessment (2025), https://arxiv.org/abs/2403.11176
3. Ali-bey, A., Chaib-draa, B., Giguère, P.: MixVPR: Feature mixing for visual place recognition. In: CVPR. pp. 14254–14263 (2023)
4. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR. pp. 3674–3683 (2018)
5. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: CVPR. pp. 5297–5307 (2016)
6. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5664–5673 (2019)
7. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
8. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: CVPR. pp. 4878–4888 (2022)
9. Brachmann, E., Cavallari, T., Prisacariu, V.A.: Accelerated coordinate encoding: Learning to relocalize in minutes using RGB and poses. In: CVPR. pp. 5044–5053 (2023)
10. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: DSAC – differentiable RANSAC for camera localization. In: CVPR. pp. 6684–6692 (2017)
11. Brachmann, E., Rother, C.: Learning less is more – 6D camera localization via 3D surface regression. In: CVPR. pp. 4654–4662 (2018)
12. Brahmbhatt, S., Gu, J., Kim, K., Hays, J., Kautz, J.: Geometry-aware learning of maps for camera localization. In: CVPR. pp. 2616–2625 (2018)
13. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. 16th European Conference on Computer Vision (ECCV) (2020)
14. Chen, J., Barath, D., Armeni, I., Pollefeys, M., Blum, H.: "where am i?" scene retrieval with language. In: European Conference on Computer Vision. pp. 201–220. Springer (2024)
15. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2017)
16. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-supervised interest point detection and description. In: CVPRW. pp. 224–236 (2018)
17. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., Gan, C., de Melo, C., Tenenbaum, J.B., Torralba, A., Shkurti, F., Paull, L.: ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In: IEEE Int. Conf. Robot. Automat. (ICRA) (2024)

18. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition. In: CVPR. pp. 14141–14152 (2021)
19. Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., Gould, S.: VLN-BERT: A recurrent vision-and-language BERT for navigation. In: CVPR. pp. 1643–1653 (2021)
20. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3D-LLM: Injecting the 3D world into large language models. In: NeurIPS (2023)
21. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In: ICCV. pp. 2938–2946 (2015)
22. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: LERF: Language embedded radiance fields. In: ICCV. pp. 19729–19739 (2023)
23. Kolmet, M., Zhou, Q., Osep, A., Leal-Taixé, L.: Text2Pos: Text-to-point-cloud cross-modal localization. In: CVPR. pp. 15172–15181 (2022)
24. Kulesza, A., Taskar, B.: Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning **5**(2-3), 123–286 (2012)
25. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: SQA3D: Situated question answering in 3D scenes. In: ICLR (2023)
26. Miao, Y., Engelmann, F., Vysotska, O., Tombari, F., Pollefeys, M., Baráth, D.B.: Scenegraphloc: Cross-modal coarse visual localization on 3d scene graphs. In: European Conference on Computer Vision. pp. 127–150. Springer (2024)
27. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
28. Peng, S., Genova, K., Jiang, C.M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.: OpenScene: 3D scene understanding with open vocabularies. In: CVPR. pp. 815–824 (2023)
29. Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., van den Hengel, A.: REVERIE: Remote embodied visual referring expression in real indoor environments. In: CVPR. pp. 9982–9991 (2020)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
31. Sarkar, S.D., Miksik, O., Pollefeys, M., Barath, D., Armeni, I.: Sgaligner: 3d scene alignment with scene graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21927–21937 (2023)
32. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: CVPR. pp. 12716–12725 (2019)
33. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR. pp. 4938–4947 (2020)
34. Sarlin, P.E., DeTone, D., Yang, T.Y., Avetisyan, A., Straub, J., Malisiewicz, T., Rota Bulo, S., Newcombe, R., Kontschieder, P., Balntas, V.: OrienterNet: Visual localization in 2D public maps with neural matching. In: CVPR. pp. 21632–21642 (2023)
35. Sarlin, P.E., Unagar, A., Lärsson, M., Germain, H., Toft, C., Larsson, V., Pollefeys, M., Lepetit, V., Kahl, L., Sattler, T.: Back to the feature: Learning robust camera localization from pixels to pose. In: CVPR. pp. 3247–3257 (2021)
36. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & effective prioritized matching for large-scale image-based localization. IEEE Trans. Pattern Anal. Mach. Intell. **39**(9), 1744–1756 (2017)

37. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: InLoc: Indoor visual localization with dense matching and view synthesis. In: CVPR. pp. 7199–7208 (2018)
38. Thomason, J., Murray, M., Cakmak, M., Zettlemoyer, L.: Vision-and-dialog navigation. In: Kaelbling, L.P., Kragic, D., Sugiura, K. (eds.) Proceedings of the Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 100, pp. 394–406. PMLR (30 Oct–01 Nov 2020), https://proceedings.mlr.press/v100/thomason20a.html
39. Wald, J., Avetisyan, A., Navab, N., Tombari, F., Nießner, M.: Rio: 3d object instance re-localization in changing indoor environments. In: Proceedings IEEE International Conference on Computer Vision (ICCV) (2019)
40. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
41. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
42. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: SceneGraphFusion: Incremental 3D scene graph prediction from RGB-D sequences. In: CVPR. pp. 7515–7525 (2021)
43. Zhang, C., Li, M., Budvytis, I., Liwicki, S.: Dialoc: An iterative approach to embodied dialog localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12585–12593 (2024)
44. Zhou, G., Hong, Y., Wu, Q.: NavGPT: Explicit reasoning in vision-and-language navigation with large language models. In: AAAI. pp. 7641–7649 (2024)
45. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3d-vista: Pre-trained transformer for 3d vision and text alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2911–2921 (2023)