

# Learning to Evolve: Multi-modal Interactive Fields for Robust Humanoid Navigation in Dynamic Environments

Peifeng Jiang<sup>1</sup>, Hong Liu<sup>1\*</sup>, Jin Jin<sup>2</sup>, Wenshuai Wang<sup>1</sup>, Xia Li<sup>3</sup>

<sup>1</sup> State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School

<sup>2</sup> Oxford Robotics Institute, University of Oxford

<sup>3</sup> Institute for Machine Learning, Department of Computer Science, ETH Zurich

jpjf@stu.pku.edu.cn, hongliu@pku.edu.cn, jinjin@robots.ox.ac.uk,  
wangws@stu.pku.edu.cn, xia.li@inf.ethz.ch

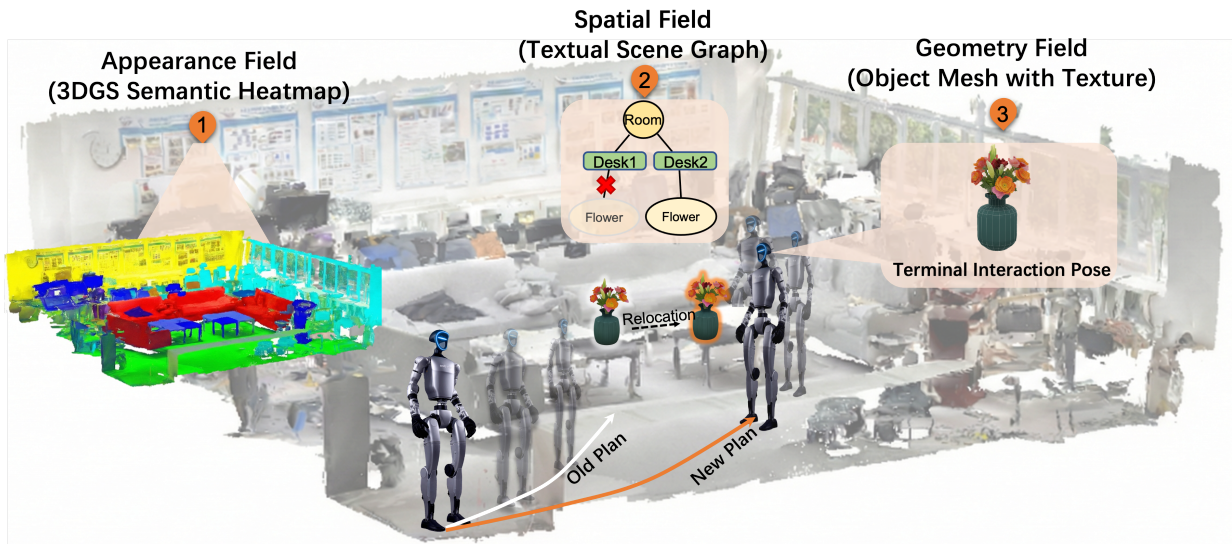


Fig. 1: **Multi-modal Interactive Fields (MIF) for Robust Humanoid Navigation.** We propose a hierarchical framework composed of three coupled fields: (1) **Appearance Field:** Provides dense semantic grounding for robust view synthesis. (2) **Spatial Field:** Maintains a dynamic topological Scene Graph that updates upon object relocation, enabling seamless re-planning from obsolete paths to correct trajectories. (3) **Geometry Field:** Generates water-tight meshes to ensure Interaction Pose Safety (IPS) before manipulation.

**Abstract**—Safe manipulation-oriented navigation for humanoid robots requires scene memory that remains reliable under locomotion-induced perceptual distortion, environmental changes, and interaction-level geometric safety constraints. Existing semantic mapping and scene-graph systems are difficult to deploy directly in this setting because they often assume stable camera trajectories, static environments, or coarse object geometry. We introduce the Multi-modal Interactive Field (MIF), a humanoid-oriented system that integrates confidence-aware semantic 3D Gaussian Splatting, discrepancy-triggered spatial memory updates, and task-driven geometric reconstruction within a closed-loop perception-adaptation pipeline. MIF couples three fields: an uncertainty-aware 3DGS Appearance Field that suppresses gait-induced blur, a Spatial Field that maintains topological memory, and a Geometry Field that supports Interaction Pose Safety (IPS) before manipulation. A discrepancy

detection score is introduced to separate locomotion-induced false-positive changes from persistent changes and updates only locally inconsistent regions. On a Unitree-G1 humanoid in a real dynamic office, MIF improves relocation success in non-static environments from 12% to 94% compared with static scene-graph memory, while reducing semantic memory footprint by 91.4% through feature distillation for practical online operation. Project page and code: <https://ziya-jiang.github.io/MIF-homepage/>.

**Index Terms**—Semantic Scene Understanding, Safe Navigation, Humanoid Robots, Active Perception, 3D Gaussian Splatting, Multiplex Interactive Learning

## I. INTRODUCTION

Humanoid robots are increasingly expected to operate in human-centric environments where navigation is tightly

coupled with subsequent physical interaction. Approximate navigation to an object is therefore insufficient: the robot must reach a terminal stance and body configuration from which manipulation is collision-free, kinematically reachable, and stable. This requires scene memory that grounds open-vocabulary semantics into reliable geometry while remaining robust to locomotion-induced perceptual distortion and environmental changes.

Unlike wheeled platforms, humanoids introduce bipedal dynamics, including torso oscillations and abrupt viewpoint shifts, that can degrade egocentric RGB-D observations during mapping [1, 2]. This challenges a common assumption behind many semantic mapping and scene-graph pipelines, such as HOV-SG [3] and ConceptGraphs [4], which are typically built from stable, quasi-static observations. At the same time, such systems often follow a map-once-query-forever paradigm, whereas human environments contain relocated, removed, or newly introduced objects. Together, these assumptions expose two coupled failure modes for manipulation-oriented humanoid navigation:

- **P1: Locomotion-Induced Semantic-Geometric Distortion.** Humanoid locomotion induces camera jitter and noisy state estimation for pose that can cause unstable observations to be fused into the map. This produces blurred semantics and noisy geometry, degrading both object grounding and the surface evidence needed to compute a safe terminal interaction pose.
- **P2: Map-Reality Mismatch under Scene Changes.** Real-world layouts change as objects are relocated, removed, or added, while static scene memories may continue to guide the robot toward obsolete coordinates. For a humanoid, stale memory not merely cause error in semantic grounding: it can lead to unsafe interaction poses, unnecessary re-planning, or failures to recognize task-relevant changes.

To address these coupled failures, we propose the Multi-modal Interactive Field (MIF), a humanoid-oriented system that maintains appearance, spatial, and geometry fields within a closed-loop perception-adaptation pipeline, as shown in Fig. 1.

For P1, the Appearance Field builds a confidence-aware semantic 3D Gaussian Splatting representation in which each Gaussian carries a reliability estimate for identifying gait-corrupted primitives. This reliability gate suppresses gait-corrupted primitives during rendering and scene-graph construction, reducing the propagation of locomotion artifacts into semantic grounding. For interaction-level safety, the Geometry Field selects better views around the target object and uses a Flow Matching-based model [5, 6] to recover object-centric meshes on demand. The recovered mesh provides the geometric proxy required by Interaction Pose Safety (IPS), allowing navigation completion to depend on collision-awareness, kinematically reachability, and stability rather than just coordinate-level arrival.

For P2, MIF uses a closed-loop Interaction and Adaptation Mechanism that treats scene memory as locally revisable dur-

ing task execution. Rather than globally re-scanning the environment, the system monitors a multi-modal discrepancy score  $\mathcal{D}$  between the live local graph and the stored global graph. Because  $\mathcal{D}$  is gated by reliability and aggregates positional, semantic, and relational evidence, it separates locomotion-induced false-positive changes from persistent scene changes. When  $\mathcal{D}$  exceeds the selected threshold, MIF triggers a spatially selective update that revises only the inconsistent region. This local evolution loop avoids the cost of global re-mapping while supporting adaptation to relocated, removed, and newly added objects. Accordingly, our contributions emphasize the integrated system:

- MIF provides a humanoid-oriented integration of confidence-aware semantic 3D Gaussian Splatting, VLM-based scene-graph reasoning, and object-centric geometry recovery for manipulation-safety-aware navigation.
- A discrepancy-triggered local update mechanism compares live and stored graph evidence, filters locomotion-induced false positives through confidence gating, and revises only inconsistent regions of the spatial memory.
- Interaction Pose Safety (IPS) is formulated as a system-level navigation criterion, supported by target-centered view selection and on-demand Flow Matching-based mesh recovery.
- Real-world validation on a Unitree G1 humanoid in a dynamic office reports task success, memory, and latency evidence for online deployment. Under non-static environment settings, MIF improves relocation success from 12% to 94% compared with static scene-graph memory.

## II. RELATED WORK

### A. Semantic 3D Scene Representations

Neural Radiance Fields (NeRF) [7] and Instant NGP [8] have advanced continuous 3D scene modeling. 3D Gaussian Splatting (3DGS) [9] further offers explicit primitives and real-time rendering, making it attractive for robotic mapping and online scene representations. To attach open-vocabulary semantics to such representations, prior work [10, 11, 12, 13, 14, 15, 16] integrates vision foundation models [11, 17, 18] into 3D fields. Extensions such as Gaussian Grouping [19], LangSplat [20], and Feature Splatting [21] distill or attach high-dimensional semantic features to Gaussian primitives for object-centric understanding, language querying, and editing. LangSplat reduces language-feature storage through a scene-wise autoencoder, while LatentBKI [22] studies open-vocabulary mapping in visual-language latent spaces with quantifiable uncertainty. These methods reduce the cost of open-vocabulary 3D querying, but dense semantic embeddings can still impose nontrivial memory and latency overhead for online humanoid deployment.

In parallel, motion-deblurring 3DGS methods such as Deblur-GS [23] reconstruct sharper Gaussian radiance fields from camera motion-blurred images, while uncertainty-aware methods such as PUP 3D-GS [24] estimate Gaussian sensitivity for pruning and efficient rendering. However, these lines of

work are usually studied as reconstruction, feature storage, or pruning problems, rather than as a coupled humanoid system that must use the reliability signal for image rendering, scene-graph construction, and manipulation-safety verification. For humanoid deployment, bipedal locomotion can introduce gait-induced artifacts, while dense semantic features can also create memory and latency costs.

MIF addresses these constraints by integrating confidence-gated semantic 3DGS with feature distillation: the former suppresses gait-corrupted primitives for P1, while the latter is an efficiency-enabling design for practical online deployment rather than a standalone novelty claim.

### B. 3D Scene Graphs and Hierarchical Mapping

3D Scene Graphs (3DSGs) provide a structured abstraction by organizing environments into floors, rooms, objects, and their relations [25, 3, 26, 27, 28]. Early systems such as Kimera-Semantics [29] focused on geometric-semantic consistency, while Hydra [30] introduced real-time spatial perception. With the advent of large language and vision-language models [31], recent frameworks [3, 32, 33, 34, 35] leverage symbolic or language-conditioned reasoning for complex tasks.

Dynamic scene-graph and spatio-temporal mapping methods, such as Khronos [36], further consider how scene memories evolve when objects move, disappear, or appear. ConceptGraphs [4] can also be adapted with periodic rescanning, providing a useful rescan-style baseline for testing whether scene memories can be refreshed after changes. Nevertheless, static scene-graph memories expose a map-once-query-forever failure mode, while dynamic or rescan-style alternatives may rely on revisiting, periodic rescanning, or sensing assumptions that differ from our non-revisiting humanoid navigation protocol. Moreover, graph nodes and coarse object boxes often lack the surface-level geometry needed to evaluate manipulation-ready terminal poses.

MIF bridges this gap by coupling the Spatial Field with an on-demand Geometry Field, while using discrepancy-triggered updates to revise only inconsistent regions of the memory during navigation.

### C. Generative Geometry for Embodied AI

Robotic interaction requires reliable geometric priors, especially when only partial observations are available. Learning-based mesh generation has been studied through methods such as BSP-based reconstruction, Scan2Mesh, AtlasNet-style surface generation, and PolyGen [37, 38, 39, 40]. Generative models, including diffusion-based [41] and Flow Matching-based [6] approaches, have further improved shape recovery from sparse or partial observations. Recent studies such as MeshGPT [42] and SAM-3D [5] explore object-centric completion that can support downstream manipulation. However, these generative priors are typically used as offline or object-centric modules, rather than being integrated with online navigation and memory adaptation.

MIF incorporates a Flow Matching-based Geometry Field that recovers object-centric meshes *on demand* during the approaching phase and uses them as geometric proxies for Interaction Pose Safety (IPS) verification. Our use of generative geometry is therefore not a standalone mesh-generation contribution; rather, it provides an on-demand geometric proxy selected from target-centered views and used by IPS during humanoid navigation.

### D. Active Perception and Interactive Learning

Interactive perception [43] studies how an agent’s actions can reduce perceptual ambiguity. Related work spans low-level active SLAM [44, 45, 46, 47, 48] and high-level semantic exploration [49, 50, 51, 52]. Modern foundation models provide useful zero-shot priors, but they can still hallucinate or produce stale semantic assumptions in novel environments [53, 51]. Although active perception connects sensing with action, humanoid-specific constraints such as locomotion-induced viewpoint oscillations remain less explored. In contrast to coverage-oriented exploration, MIF uses two task-driven mechanisms: target-centered view selection for interaction geometry and discrepancy-triggered local memory revision for map-reality conflicts. This makes scene memory locally revisable during navigation, moving the humanoid beyond passive use of a static map.

## III. METHODOLOGY

We propose the Multi-modal Interactive Field (MIF), a humanoid-oriented system representation  $\mathcal{M} = \{\mathcal{F}_{app}, \mathcal{F}_{spat}, \mathcal{F}_{geom}\}$  for safe manipulation-oriented navigation from egocentric RGB-D observations. Unlike static mapping pipelines, MIF is designed to mitigate *locomotion-induced distortion* (P1) and *map-reality mismatches* (P2) by grounding language queries into spatial memory and interaction-ready geometry. The pipeline operates as a persistent perception-adaptation loop that incrementally stabilizes appearance, updates spatial memory, and verifies interaction geometry. The overview of the Multi-modal Interactive Fields (MIF) framework is shown in Fig 2. In the following sections, we detail: (i) the construction of the confidence-gated Appearance Field  $\mathcal{F}_{app}$  via jitter-aware semantic 3DGS (Sec. III-A); (ii) the abstraction of the Spatial Field  $\mathcal{F}_{spat}$  from stabilized rendered views for semantic grounding and mismatch detection (Sec. III-B); and (iii) the on-demand generation of the Geometry Field  $\mathcal{F}_{geom}$  for Interaction Pose Safety (IPS) verification (Sec. III-C).

### A. Appearance Field Construction ( $\mathcal{F}_{app}$ )

The Appearance Field  $\mathcal{F}_{app}$  serves as the sensory grounding layer of MIF. Unlike standard reconstruction pipelines that assume relatively stable camera motion,  $\mathcal{F}_{app}$  is designed to mitigate the semantic-geometric distortion (P1) caused by bipedal locomotion. Rather than directly fusing all raw frames, it uses per-Gaussian reliability to downweight observations likely affected by high-frequency torso oscillations and local



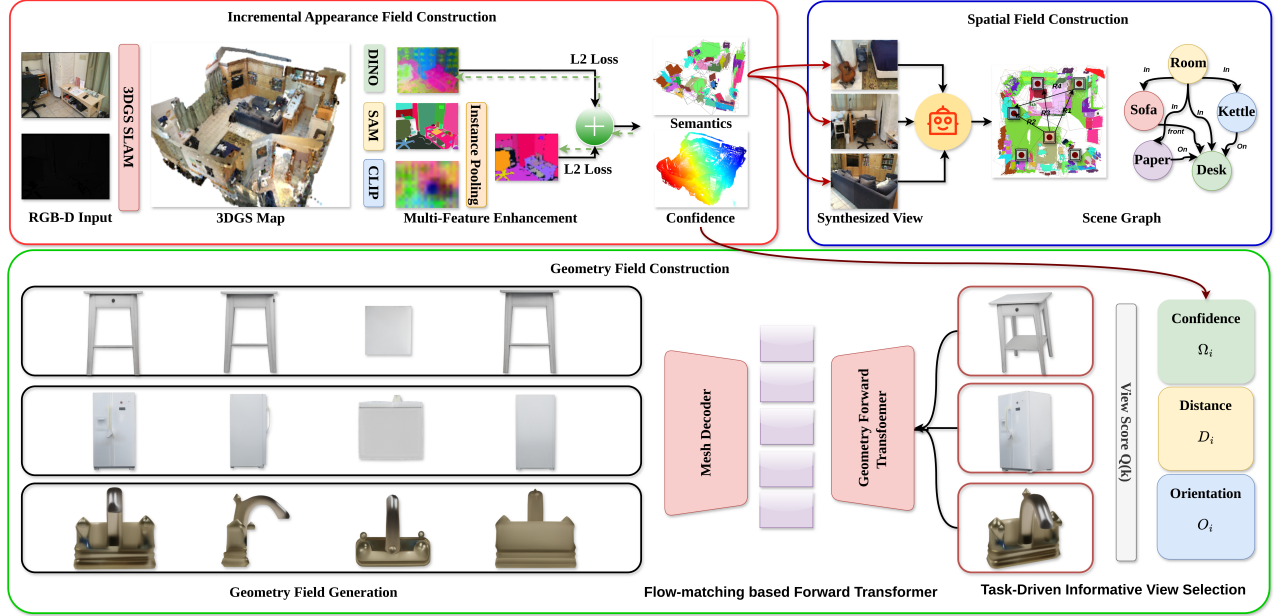


Fig. 2: **Overview of the Multi-modal Interactive Fields (MIF) framework.** (Red) Incremental Appearance Field: Constructs a dense semantic base by fusing 3DGS SLAM [54] with multi-modal features [18, 17] to generate confidence-aware maps. (Blue) Spatial Field: Abstracts the dense map into a topological Scene Graph, utilizing VLMs [31] to reason over object relationships from synthesized views. (Green) Geometry Field: Recovers high-fidelity meshes for interaction. It employs a Task-Driven View Selection strategy for object-centric geometry recovery by adapting a Flow-matching Transformer [5].

pose uncertainty, producing a more view-consistent semantic radiance field.

We represent the field as a collection of 3D Gaussian primitives  $\mathcal{G} = \{G_i\}$ . Each primitive is augmented with a confidence scalar  $C_i \in [0, 1]$ , which acts as a reliability gate for identifying and suppressing gait-corrupted primitives.

1) *Jitter-Aware Confidence Estimation*: To reduce transient artifacts such as motion blur and pose-induced misalignment, we introduce a differentiable confidence module. The key observation is that gait-corrupted primitives tend to remain optimization-unstable across training, whereas stable high-texture boundaries usually converge after multi-view optimization. Accordingly, we compute  $C_i$  from each Gaussian’s weighted gradient magnitude and opacity, normalized by their global means as  $g_{n,i}$  and  $\alpha_{n,i}$ , respectively. The subscript  $n$  denotes normalization by the corresponding global mean. The gradient term aggregates instability from position, scale/rotation, color, and opacity updates, while opacity reflects whether a primitive has accumulated stable structural evidence.

The confidence  $C_i$  is then computed through a smooth gating mechanism that penalizes high instability while preserving structurally supported primitives:

$$C_i = \exp(-\beta g_{n,i}) + (1 - \exp(-\beta g_{n,i})) \cdot \sigma(\gamma \alpha_{n,i}(1 - g_{n,i})), \quad (1)$$

where  $\sigma(\cdot)$  denotes the sigmoid function, which maps the term  $\gamma \alpha_{n,i}(1 - g_{n,i})$  into the range  $[0, 1]$ .  $\beta$  controls the penalty on optimization instability and  $\gamma$  controls the opacity-based reliability scaling. This mechanism downweights low-

confidence, ghosting-like artifacts in rendering and object reliability estimation, reducing their influence on the subsequent spatial graph.

2) *Open-Vocabulary Feature Distillation*: To enable online semantic querying with bounded memory, we distill high-dimensional CLIP/DINOv2 features [17, 18] into a compact 32-dimensional latent space  $\mathbf{f}_i^{distill} \in \mathbb{R}^{32}$ . This compression reduces the memory footprint needed to maintain large-scale semantic maps and supports practical online deployment (see Sec. IV-E).

A co-trained MLP decoder projects the latent features back to the original CLIP/DINOv2 feature space for open-vocabulary matching. Feature rendering follows standard alpha-blending but is weighted by the confidence metric so that unstable primitives contribute less to semantic evidence:

$$\mathbf{F}^{distill}(\mathbf{p}) = \sum_{i \in \mathcal{N}} C_i \mathbf{f}_i^{distill} \alpha_i T_i, \quad \text{where } T_i = \prod_{j=1}^{i-1} (1 - C_j \alpha_j). \quad (2)$$

Here,  $\mathbf{F}^{distill}(\mathbf{p})$  denotes the rendered distilled semantic feature at pixel  $\mathbf{p}$ , obtained by accumulating per-primitive features along the ray. By coupling feature accumulation with  $C_i$ , the field reduces semantic smearing (P1) and anchors semantic evidence primarily to stable regions.

3) *Stabilized View Synthesis for Spatial Reasoning*: Once optimized,  $\mathcal{F}_{app}$  functions as a confidence-weighted rendering engine for downstream spatial reasoning. Instead of feeding raw, gait-affected camera frames to the perception stack, we render stabilized views  $I_{stable}$  from smoothed virtual trajectory-



ries. These rendered views reduce gait-induced artifacts before object detection and the extracting spatial relation, providing more stable evidence for subsequent Spatial Field construction. A later discrepancy-computation step uses the node positions, semantic features, and relation sets extracted from these views to determine whether a local memory update is needed.

### B. Spatial Field Construction ( $\mathcal{F}_{spat}$ )

The Spatial Field  $\mathcal{F}_{spat}$  serves as the topological memory and reasoning layer of MIF. While  $\mathcal{F}_{app}$  provides confidence-weighted visual evidence,  $\mathcal{F}_{spat}$  abstracts the scene into a hierarchical graph  $\mathcal{M}_{spat} = (\mathcal{V}, \mathcal{E})$  for high-level task reasoning. Its function is twofold: (1) grounding linguistic concepts into physical space using stabilized evidence, and (2) detecting likely Map-Reality Mismatches (P2) by comparing local and stored graph structure.

1) *Graph Generation via Stabilized Re-perception*: Direct object detection and semantic parsing on raw humanoid camera streams can suffer from segmentation jitter under gait-induced viewpoint changes. We therefore employ a stabilized re-perception strategy. We query the optimized  $\mathcal{F}_{app}$  to render confidence-weighted stabilized views  $\{I_{stable}\}$ . These rendered inputs reduce gait-induced artifacts before object detection and relation extraction, providing more stable evidence for scene-graph construction. Formally, each node  $v_i \in \mathcal{V}$  is represented as  $v_i = \{\mathcal{S}_i, \Omega_i, \mathcal{A}_i, \mathbf{c}_i\}$ , containing a semantic caption  $\mathcal{S}_i$ , room affiliation  $\mathcal{A}_i$ , node reliability  $\Omega_i$ , and 3D centroid  $\mathbf{c}_i$  computed via confidence-weighted back-projection:

$$\mathbf{c}_i = \frac{\sum_{j \in \mathcal{N}_i} C_j \alpha_j \mathbf{x}_j}{\sum_{j \in \mathcal{N}_i} C_j \alpha_j}, \quad (3)$$

where  $\mathbf{x}_j$  denotes the mean position of the  $j$ -th Gaussian primitive associated with node  $v_i$ . We define the node-level reliability score as  $\Omega_i = \langle C_j \rangle_{j \in \mathcal{N}_i}$ , where  $\langle \cdot \rangle$  denotes the arithmetic mean over associated Gaussians. This score acts as a reliability filter: detections supported mainly by low-confidence primitives receive low  $\Omega_i$ , reducing the chance that gait-induced artifacts or unstable semantic proposals become persistent graph nodes.

2) *Detecting Map-Reality Mismatch via Discrepancy  $\mathcal{D}$* : To address P2, the system must separate transient sensing artifacts from persistent environmental changes. We implement a closed-loop Interaction and Adaptation Mechanism that monitors a multi-modal discrepancy score  $\mathcal{D}$ . The agent compares a live local subgraph  $\mathcal{M}_{loc}$  against the stored global graph  $\mathcal{M}_{spat}$ . For a matched local-global node pair, the node-level discrepancy  $\delta_i$  is defined as:

$$\delta_i = \Omega_i \cdot \left( w_{pos} \|\mathbf{c}_i^{loc} - \mathbf{c}_j^{glob}\|_2 + w_{sem} (1 - \cos(\hat{\mathbf{f}}_i, \hat{\mathbf{f}}_j)) \right). \quad (4)$$

Here,  $\mathbf{c}_i^{loc}$  and  $\mathbf{c}_j^{glob}$  are matched local/global centroids,  $\hat{\mathbf{f}}_i$  and  $\hat{\mathbf{f}}_j$  are normalized semantic features,  $w_{pos}$  and  $w_{sem}$  weight geometric and semantic drift. This reduces the influence of

low-confidence nodes on change detection, limiting false positives caused by motion blur or gait-induced artifacts (P1). The total discrepancy  $\mathcal{D}$  aggregates node-level shifts and relational changes:

$$\mathcal{D} = \frac{1}{|\mathcal{V}_{loc}|} \sum_{i \in \mathcal{V}_{loc}} \delta_i + w_{rel} \frac{|\mathcal{E}_{loc} \Delta \mathcal{E}_{spat}|}{|\mathcal{E}_{loc} \cup \mathcal{E}_{spat}|}. \quad (5)$$

Here,  $\mathcal{V}_{loc}$  averages local node discrepancies, while  $\mathcal{E}_{loc}$  and  $\mathcal{E}_{spat}$  are live and stored edge sets;  $\Delta$  and  $|\cdot|$  denote symmetric difference and cardinality, with  $w_{rel}$  weighting relational inconsistency. When  $\mathcal{D} > \tau$ , the system treats the local region as a likely Map-Reality Mismatch and selectively updates only the inconsistent region using accumulated multi-view evidence. Implementation-level definitions of the three fields and the partial graph-update procedure are provided in Sec. S.3 and Algorithm 1.

### C. Geometry Field Construction ( $\mathcal{F}_{geom}$ )

The Geometry Field  $\mathcal{F}_{geom}$  serves as the interaction-safety layer of MIF. Although  $\mathcal{F}_{app}$  and  $\mathcal{F}_{spat}$  enable semantic grounding and topological reasoning, their sparse or implicit geometry is insufficient for selecting terminal poses for manipulation. Safe humanoid manipulation benefits from explicit object surfaces that can support collision and reachability checks. To bridge this gap,  $\mathcal{F}_{geom}$  functions as an on-demand module that recovers object-centric meshes from stable visual evidence and uses them to evaluate Interaction Pose Safety (IPS).

1) *Task-Local View Selection for Interaction Geometry*: To reduce geometric ambiguity before interaction, the agent employs a task-local active gaze strategy around the target object. This process samples candidate viewpoints in the local neighborhood of the object on a local hemisphere and ranks them by a geometric observation utility for the object's interaction surfaces. We rank candidate viewpoints using the task-local utility  $Q(k)$ :

$$Q(k) = \exp(-\|\mathbf{c}_i - \mathbf{p}_k\|_2^2 / 2\sigma_d^2) \cdot (\mathbf{d}_k \cdot \mathbf{v}_{ik})^\gamma \cdot \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \Omega_j, \quad (6)$$

where  $\mathbf{p}_k$  and  $\mathbf{d}_k$  denote the position and optical-axis direction of candidate view  $k$ , and  $\mathbf{v}_{ik} = (\mathbf{c}_i - \mathbf{p}_k) / \|\mathbf{c}_i - \mathbf{p}_k\|_2$  is the normalized direction from the camera center to the object centroid  $\mathbf{c}_i$ . The distance term favors nearby views, while  $(\mathbf{d}_k \cdot \mathbf{v}_{ik})^\gamma$  favors foveal alignment between the optical axis and the object center. The confidence term favors views that observe high-reliability Gaussian neighborhoods, so the generative model is conditioned on stable evidence rather than gait-corrupted observations. This links active perception to interaction feasibility by improving the observability of surfaces that are relevant to the intended manipulation.

2) *Conditional Flow Matching for 3D Mesh Recovery*: To infer partially observed object surfaces (e.g., the occluded back side), we employ a Flow Matching-based [6] reconstruction model [5]. We formulate geometric recovery as a

conditional generation problem. Specifically, the model learns a time-dependent vector field that transports a Gaussian noise distribution toward the target mesh distribution conditioned on the selected observation.

This generation is conditioned on the selected view ranked by the task-local utility  $Q(k)$ . By conditioning on this high-confidence visual evidence, the model recovers an object-centric mesh that is consistent with the denoised observations from  $\mathcal{F}_{app}$  and can be used by the IPS check.

3) *Asynchronous Grounding and IPS Verification*: To avoid blocking navigation during generative inference, we implement an asynchronous grounding pipeline. While the robot approaches the target, the selected-view mesh is synthesized asynchronously and aligned to the Gaussian support using scale-invariant ICP:

$$\min_{\mathbf{R}, \mathbf{t}, s} \sum_{m=1}^M \rho \left( \|s\mathbf{R}\mathbf{x}_{gen}^m + \mathbf{t} - \mathbf{x}_{app}^{NN(m)}\| \right). \quad (7)$$

Here, ICP estimates the rotation  $\mathbf{R}$ , translation  $\mathbf{t}$ , and scale  $s$  that align each generated mesh point  $\mathbf{x}_{gen}^m$  to its nearest Gaussian-support point  $\mathbf{x}_{app}^{NN(m)}$ , with robust loss  $\rho$  reducing outlier influence. The registered mesh  $\mathcal{M}_{gen}$  provides dense local geometry for evaluating the final standing pose  $\mathbf{p}_{se3}$ .

4) *Interaction Pose Safety*: The IPS score  $S_{IPS}(\mathbf{p}_{se3})$  is an operational safety criterion for checking whether the terminal pose is suitable for manipulation:

$$S_{IPS} = \mathbf{I}_{col}(\mathbf{p}_{se3}, \mathcal{M}_{gen}) \wedge \mathbf{I}_{ik}(\mathbf{p}_{se3}, \mathbf{t}_{obj}) \wedge \mathbf{I}_{stab}(\mathbf{p}_{se3}), \quad (8)$$

where:

- $\mathbf{I}_{col}$  ( Geometric Safety): evaluates the signed distance to  $\mathcal{M}_{gen}$  and checks whether the collision margin exceeds  $\delta_{safe}$ .
- $\mathbf{I}_{ik}$  ( Kinematic Reachability): returns 1 if a valid collision-free inverse-kinematics solution exists for reaching the target point  $\mathbf{t}_{obj}$ .
- $\mathbf{I}_{stab}$  ( Postural Stability): returns 1 if the projected center of mass (CoM) lies within the support polygon  $\mathcal{P}_{sup}$ , defined by the foot contact region at pose  $\mathbf{p}_{se3}$ .

Thus, navigation is considered complete only when the robot reaches a pose that is collision-free, kinematically reachable, and stable for the subsequent manipulation task, rather than merely arriving near the target coordinate.

#### D. Language-Conditioned Navigation and Adaptation

The final stage converts language-conditioned goals into navigation and verification actions. Using the hierarchical structure of MIF, the system selects semantically grounded targets, verifies geometry-aware terminal poses, and updates local memory when persistent scene changes are detected.

1) *Hierarchical Semantic Grounding*: Following a hierarchical search strategy similar to [3], the agent parses a natural-language instruction (e.g., “Grasp the blue mug on the table”) into a  $\langle \text{Region, Landmark, Object} \rangle$  triplet using a VLM [31].

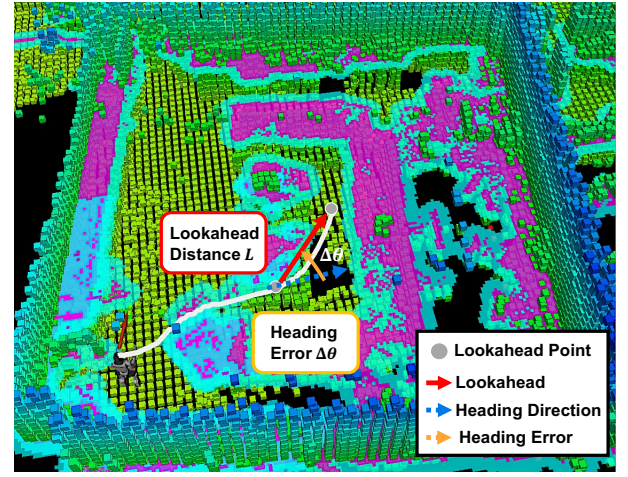


Fig. 3: **Humanoid navigation control via Pure Pursuit.** The schematic illustrates the geometric tracking logic where the Unitree-G1 robot targets a lookahead point at distance  $L$  along the planned trajectory. The heading error  $\Delta\theta$  is used to compute the required curvature  $\kappa = 2\sin(\Delta\theta)/L$ . The overlaid velocity curve demonstrates the adaptive scaling mechanism, where  $L$  is dynamically adjusted to suppress oscillations and maintain humanoid balance during high-curvature maneuvers.

The system sequentially queries  $\mathcal{F}_{spat}$  to constrain the search space by localizing the room, identifying the landmark, and retrieving the target object node  $v_i$ . The object’s centroid  $\mathbf{c}_i$ , computed by Eq. 3, provides the initial navigation goal, while the final stance is verified by IPS before interaction.

2) *Stability-Aware Execution and Safety Verification*: To manage humanoid locomotion constraints, we use a stability-aware local planner that prioritizes center-of-mass (CoM) stability when tracking the planned path. The planner scales velocity according to angular tracking error to reduce gait-induced oscillations and preserve perception quality during motion, as shown in Fig. 3.

Navigation completion is defined by the IPS criterion rather than simple coordinate arrival. Before manipulation, the agent evaluates the candidate stance against the registered geometry  $\mathcal{M}_{gen}$ . If the IPS checks indicate collision risk or insufficient reachability, the local planner adjusts the stance within the nearby feasible region and re-evaluates  $S_{IPS}$ .

3) *Closed-Loop Memory Evolution*: During navigation, the Interaction and Adaptation Mechanism uses new observations to validate the local consistency of the stored memory. The discrepancy score  $\mathcal{D}$  distinguishes between two operational cases:

- **Transient Discrepancy**: temporary obstacles (e.g., moving people) trigger local avoidance without updating the persistent scene memory.
- **Structural Mismatch**: a persistent increase in  $\mathcal{D}$  above  $\tau$  indicates a likely map-reality conflict (e.g., relocated furniture). The system then scans the discrepant region, updates only the locally inconsistent parts of  $\mathcal{F}_{app}$  and  $\mathcal{F}_{spat}$ , and re-grounds the query in the updated topology.

#### IV. EXPERIMENTAL EVALUATION

The experiments evaluate whether MIF enables a humanoid robot to maintain, query, and revise a multi-modal scene memory during real-world navigation and interaction. We organize the evaluation around the two vulnerabilities introduced earlier: locomotion-induced semantic-geometric distortion (P1) and map-reality mismatch under scene changes (P2). The main paper reports the following three questions, with additional implementation details and visualizations provided in the supplementary material.

- **Q1 (Solving P1):** Can the Appearance and Spatial Fields preserve semantic grounding under bipedal camera jitter and fast walking?
- **Q2 (Enabling Interaction):** Does target-centered geometry reconstruction improve IPS-verified interaction feasibility compared with point-cloud and static-fusion alternatives?
- **Q3 (Solving P2):** Can the discrepancy-triggered adaptation loop distinguish persistent scene changes from transient sensing artifacts and update only inconsistent local memory?

Because the system spans mapping, semantic grounding, interaction safety, and memory adaptation, we report the task-level evidence in the main paper and keep extended qualitative sequences and parameter settings in the supplementary.

##### A. Experimental Setup

We conduct real-world experiments in an approximately 100 m<sup>2</sup> office environment containing more than 100 object categories, where objects can be relocated, removed, or newly introduced after mapping. The full ROS1 [55] system runs on a centralized workstation with an NVIDIA RTX 4090 and communicates with a Unitree G1 humanoid during navigation and interaction trials. All compared methods use the same SLAM backend where applicable, so the reported differences mainly reflect memory construction, semantic grounding, geometry reconstruction, and adaptation. The supplementary material provides additional system implementation details, hyperparameter settings, and qualitative visualization results.

The baselines are selected according to each question rather than treated as a single universal set.

- For semantic-feature baselines, we include Feature-Splatting [21], LangSplat [20], LatentBKI [22], and MIF variants with different feature dimensions.
- For scene-graph and dynamic-memory baselines, we compare with HOV-SG [3], ConceptGraphs [4] with periodic rescanning, and Khronos [36] when evaluating adaptation after scene changes, without allowing a separate global rescan before task execution.
- For rendering robustness under humanoid walking, we further compare against Deblur-GS and PUP-3DGS using slow- and fast-walk image-quality metrics.

##### B. Robust Semantic Grounding under Distortion (Q1)

This experiment evaluates whether confidence-gated rendering in  $\mathcal{F}_{app}$  provides stable evidence for semantic localization

and graph construction in  $\mathcal{F}_{spat}$  under P1. The Unitree G1 receives open-vocabulary object queries, such as “Find the water bottle”, and must localize the target in the current scene memory.

1) *Quantitative Results:* Following HOV-SG, a query is counted as successful when the predicted object matches the ground-truth category and its centroid is within 1.0m of the target. Because interaction requires tighter spatial accuracy than goal-level navigation, we also report mean distance error (MDE). Table I shows that MIF improves task-level semantic localization under humanoid motion, with higher success rate and lower MDE than HOV-SG and Feature-Splatting in this setup. The lower MDE suggests that confidence-gated integration suppresses locomotion-induced smearing that otherwise shifts semantic anchors away from object centers.

TABLE I: Quantitative results of semantic grounding under locomotion-induced distortion.

| Method                 | Success Rate | MDE (m)     | Latency (s) |
|------------------------|--------------|-------------|-------------|
| Feature-Splatting [21] | 65%          | 0.42        | 3.2         |
| HOV-SG [3]             | 74%          | 0.35        | 2.8         |
| <b>MIF (Ours)</b>      | <b>92%</b>   | <b>0.18</b> | <b>1.6</b>  |

To isolate rendering robustness, Table II compares MIF with Deblur-GS and PUP-3DGS under slow and fast walking.

TABLE II: Rendering results under slow (S-) and fast (F-) walk.

| Method                             | S-PSNR↑     | S-SSIM↑     | F-PSNR↑     | F-SSIM↑     |
|------------------------------------|-------------|-------------|-------------|-------------|
| w/o Confidence                     | 28.4        | 0.88        | 21.6        | 0.72        |
| Deblur-GS [23]                     | 25.14       | 0.80        | 24.67       | 0.70        |
| PUP-3DGS [24]                      | 20.44       | 0.67        | 17.89       | 0.53        |
| <b>MIF <math>C_i</math> (Ours)</b> | <b>31.2</b> | <b>0.93</b> | <b>29.8</b> | <b>0.89</b> |

MIF is not using confidence only for visual quality; the same reliability estimate is used to gate map integration and reduce unstable evidence before graph construction.

Table III further shows that scene graphs constructed from stabilized rendered views achieve better object localization than those constructed from raw walking frames, especially under fast walking.

TABLE III: Semantic object localization on scene graphs constructed from rendered vs. raw frames.

| Input Source               | Recall↑     | Precision↑  | mAP↑        |
|----------------------------|-------------|-------------|-------------|
| Raw frames (Slow)          | 0.87        | 0.91        | 0.84        |
| Raw frames (Fast)          | 0.61        | 0.79        | 0.58        |
| <b>MIF Rendered (Slow)</b> | <b>0.94</b> | <b>0.95</b> | <b>0.93</b> |
| <b>MIF Rendered (Fast)</b> | <b>0.92</b> | <b>0.94</b> | <b>0.91</b> |
| Static camera (Ref.)       | ~1.00       | ~1.00       | ~1.00       |

##### C. High-fidelity Geometry for Interaction Safety (Q2)

This experiment evaluates the transition from object semantic grounding and navigation to IPS-verified interaction. We



test whether the Geometry Field provides a safer physical proxy for terminal interaction poses than sparse point-cloud or static-fusion alternatives.

1) *The Necessity of High-fidelity Geometry*: Compared with raw point-cloud planning, MIF reduces IPS failure from 38% to 6% in our trials, mainly because continuous surfaces expose feasible clearances more reliably than sparse point clouds formed by Gaussian centers. The point-cloud baseline often misses surface connectivity, while the static mesh baseline can preserve locomotion-induced artifacts that occlude otherwise feasible workspaces.

For safety, sparse point clouds can make IPS checks unreliable: without continuous surface connectivity, they provide incomplete normal and boundary information for manipulation, causing  $\mathbf{I}_{col}$  to mark some risky poses as feasible. The generated watertight meshes reduce this failure mode and make the collision term in IPS better aligned with physical contact constraints.

2) *Interaction Performance*: We count an interaction trial as successful only when the robot arrives and the terminal pose satisfies  $S_{IPS} = 1$ . MIF achieves a mean terminal position error of 0.12m in these trials. This precision supports the reachability check  $\mathbf{I}_{ik}$ , since larger terminal errors can move the target outside the feasible manipulation workspace..

The active viewpoint score  $Q(k)$  is reported as a task-local geometric observation utility: it selects views with better distance, alignment, and high-confidence Gaussian support before mesh reconstruction. These cleaner conditioning views support Flow-Matching reconstruction and improve IPS performance as part of the overall system.

3) *Baseline Comparison Analysis*: To examine these gains, we compared MIF against two ablative baselines:

- Pt-Cloud Baseline: Pt-Cloud plans on sparse Gaussian centroids; without surface connectivity, collision checks can be unreliable around gaps.
- Static Mesh Baseline: Static Mesh uses TSDF fusion without the confidence gate; locomotion artifacts can be integrated into the mesh and reduce feasible terminal-pose estimates.

As shown in Table IV, MIF attains 94% IPS success with no observed collisions, suggesting that confidence-gated mapping and on-demand geometry are complementary for interaction safety.

TABLE IV: Interaction Feasibility Comparison. We compare our generative approach against a Pt-Cloud Baseline and a Static Mesh Baseline.

| Method            | IPS Success (%) | Nav. Err. (m) | Collision Rate |
|-------------------|-----------------|---------------|----------------|
| Pt-Cloud          | 62%             | 0.28          | 32%            |
| Static Mesh       | 78%             | 0.22          | 14%            |
| <b>MIF (Ours)</b> | <b>94%</b>      | <b>0.12</b>   | <b>0%</b>      |

#### D. Dynamic Adaptation to Map-Reality Mismatches (Q3)

This experiment evaluates whether the discrepancy-triggered adaptation loop mitigates P2 after the initial map

becomes stale. After initial mapping, we introduce three types of scene changes: object relocation, removal, and addition. For relocation, the target is moved by more than 1.5 m, and the robot must adapt during task execution without a separate global rescan after the scene change.

1) *Evaluation Protocol*: A trial is successful only if the robot either reaches the target’s current physical location with an IPS-valid final pose or correctly reports that the target has been removed. Arriving at obsolete memory coordinates is counted as failure. All methods start from the same initial map and post-change task condition, and no method is allowed to perform a separate global rescan before task execution. During execution, each method uses the observations produced by its own navigation or adaptation strategy.

2) *Results Analysis*: During navigation, MIF compares the live local graph with the stored Spatial Field using  $\mathcal{D}$ , which aggregates positional, semantic, and relational evidence. Fig. 4 illustrates how a detected discrepancy triggers a local Spatial Field update after scene changes. Table V reports the ROC-based threshold selection used to set  $\tau = 0.45$ , showing the trade-off between change recall and false-positive updates.

TABLE V: ROC-based threshold selection in real scenarios.

| Threshold $\tau$   | True Positive Rate $\uparrow$ | False Positive Rate $\downarrow$ | F1 $\uparrow$ |
|--------------------|-------------------------------|----------------------------------|---------------|
| 0.25               | 99.2%                         | 18.7%                            | 0.817         |
| 0.35               | 97.8%                         | 9.3%                             | 0.893         |
| <b>0.45 (Ours)</b> | <b>95.1%</b>                  | <b>4.2%</b>                      | <b>0.934</b>  |
| 0.55               | 89.6%                         | 1.8%                             | 0.913         |
| 0.65               | 78.3%                         | 0.7%                             | 0.866         |

This threshold is fixed for all scene-change trials. Table VI then compares static-memory, periodic-rescan, dynamic scene-graph, and MIF variants under the same setting. HOV-SG has low success when the target is no longer at the stored coordinates, while ConceptGraphs+Rescan and Khronos provide stronger baselines for evaluating whether scene memories can be revised after changes. For MIF, we report two stages to separate performance without memory revision from the final performance after memory revision is enabled. MIF (Initial) disables the local memory update mechanism and only uses local observations near the obsolete memory location; it can still succeed when the relocated object remains visible or close enough for local perception and planning. MIF (Post-Update) reports the final success rate of the full system, including the initially solved cases and the additional cases recovered after  $\mathcal{D} > \tau$  triggers active scanning, local graph revision, and replanning.

#### E. System Efficiency

We finally report computational cost because online humanoid deployment requires a balanced allocation of computational resources, with sufficient headroom for perception, language grounding, geometry generation, and control.

Table VII compares MIF’s 32D and 64D variants with LangSplat, LatentBKI, and FeatSplat under slow and fast walking.

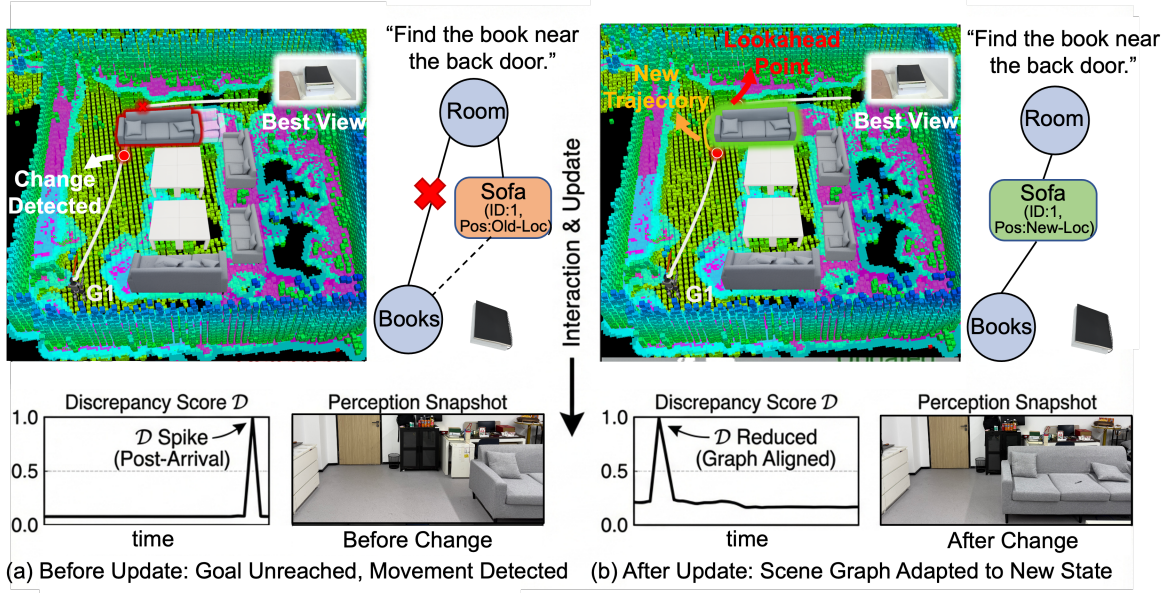


Fig. 4: **Dynamic adaptation of the hierarchical scene graph to environmental changes.** (a) **Detection:** Upon arriving at the goal, the G1 robot detects a structural discrepancy (e.g., a moved sofa) that contradicts its prior memory, triggering an increase in the discrepancy score  $\mathcal{D}$ . (b) **Resolution:** MIF initiates a local update of the Spatial Field with fresh observations. The updated graph better aligns with the physical reality, reducing  $\mathcal{D}$  and improving consistency for subsequent tasks.

TABLE VI: Scene-change task success without a separate global rescan before task execution. Reloc/Remov/Addit denote relocated ( $> 1.5\text{m}$ ), removed, and newly added objects. MIF (Initial) uses local observations near the obsolete memory location without local memory revision, while MIF (Post-Update) reports final task success after active scanning, local graph revision, and replanning are enabled.

| Method                     | Reloc SR $\uparrow$ | Remov SR $\uparrow$ | Addit SR $\uparrow$ |
|----------------------------|---------------------|---------------------|---------------------|
| HOV-SG [3]                 | 12%                 | 10%                 | 0%                  |
| ConceptGraphs + Rescan [4] | 38%                 | 35%                 | 12%                 |
| Khronos [36]               | $\sim 18\%$         | 61%                 | $\sim 12\%$         |
| MIF (Initial)              | 64%                 | 68%                 | 42%                 |
| <b>MIF (Post-Update)</b>   | <b>94%</b>          | <b>98%</b>          | <b>86%</b>          |

\*The two-session protocol first builds the initial map, then changes the scene and evaluates task completion during post-change navigation. The robot is not required to exactly repeat the initial mapping trajectory.

TABLE VII: Semantic localization under slow (S-) and fast (F-) walk conditions, memory usage, and latency.

| Method                | VRAM $\downarrow$ (GB) | Latency $\downarrow$ (s) | S-SR $\uparrow$ | F-SR $\uparrow$ |
|-----------------------|------------------------|--------------------------|-----------------|-----------------|
| LangSplat [20]        | 18.45                  | 2.3                      | 95%             | 72%             |
| LatentBKI-64D [22]    | 23.5                   | 7.1                      | <b>98%</b>      | 74%             |
| FeatSplat-32D [21]    | 4.8                    | 3.2                      | 81%             | 59%             |
| <b>MIF 64D (Ours)</b> | 18.0                   | 4.6                      | 92%             | <b>88%</b>      |
| <b>MIF 32D (Ours)</b> | <b>&lt;4.0</b>         | <b>1.6</b>               | 90%             | 86%             |

In practice,  $\mathcal{F}_{app}$  supports real-time incremental updates at 22 FPS, while the asynchronous generation of  $\mathcal{F}_{geom}$  takes around 6.2s and is scheduled during approach or verification rather than blocking low-level control. The 32D semantic feature field further reduces VRAM usage to below 4GB for

the tested office map, leaving capacity for the VLM/LLM and mesh-generation modules on the RTX 4090.

## V. CONCLUSION

We presented Multi-modal Interactive Field (MIF), a humanoid-oriented system for manipulation-safety-aware navigation in changing indoor environments. MIF integrates a confidence-aware Appearance Field, a topology-aware Spatial Field, and an object-centric Geometry Field to mitigate locomotion-induced semantic-geometric distortion (P1) and map-reality mismatch under scene changes (P2). On a Unitree G1 humanoid, MIF achieves 92% semantic-grounding success, 94% IPS success with no observed collisions, and final adaptation success rates of 94%, 98%, and 86% for relocation, removal, and addition. These results suggest that coupling reliability estimation, discrepancy-based local memory revision, and target-centered geometry verification improves humanoid interaction in non-static environments. Remaining limitations include several-second geometry generation latency and reliance on a centralized RTX 4090 workstation; future work will reduce latency, move more computation onboard, and develop multiplex interactive learning so that visual, spatial, geometric, and human feedback can jointly refine the robot’s memory and adaptation policy over repeated interactions.

## ACKNOWLEDGMENTS

This work is jointly supported by the National Natural Science Foundation of China (No.62373009), and Guangdong S&T Program (No.2024B0101050002). We specifically thank Jin Jin for his critical insights into the overall methodology and extensive revisions of the manuscript.

## REFERENCES

- [1] Olivier Stasse, Andrew J Davison, Ramzi Sellaouti, and Kazuhito Yokoi. Real-time 3d slam for humanoid robot considering pattern generator information. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 348–355, 2006.
- [2] Shuuji Kajita, Fumio Kanehiro, Kenji Kaneko, Kazuhito Yokoi, and Hirohisa Hirukawa. The 3d linear inverted pendulum mode: A simple modeling for a biped walking pattern generation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 1, pages 239–246, 2001.
- [3] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *Robotics: Science and Systems (RSS)*, 2024.
- [4] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028, 2024.
- [5] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.
- [6] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.
- [7] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [8] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics*, 41(4):1–15, 2022.
- [9] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.
- [10] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 19729–19739, 2023.
- [11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.
- [12] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–824, 2023.
- [13] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5333–5343, 2024.
- [14] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:23311–23330, 2022.
- [15] Francis Engelmann, Fabian Manhardt, Michael Niemeyer, Keisuke Tateno, and Federico Tombari. Opennerf: Open set 3d neural scene segmentation with pixel-wise features and rendered novel views. In *International Conference on Learning Representations (ICLR)*, volume 2024, pages 8396–8407, 2024.
- [16] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23171–23181, 2023.
- [17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.
- [18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024.
- [19] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European conference on computer vision (ECCV)*, pages 162–179. Springer, 2024.
- [20] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20051–20060, 2024.
- [21] Ri-Zhao Qiu, Ge Yang, Weijia Zeng, and Xiaolong Wang. Language-driven physics-based scene synthesis and editing via feature splatting. In *European conference on computer vision (ECCV)*, pages 368–383, 2024.
- [22] Joey Wilson, Ruihan Xu, Yile Sun, Parker Ewen, Minghan Zhu, Kira Barton, and Maani Ghaffari. Latentbki: Open-dictionary continuous mapping in visual



- language latent spaces with quantifiable uncertainty. *IEEE Robotics and Automation Letters*, 2025.
- [23] Jeongtaek Oh, Jaeyoung Chung, Dongwoo Lee, and Kyoungh Mu Lee. Deblurgs: Gaussian splatting for camera motion blur. *arXiv preprint arXiv:2404.11358*, 2024.
- [24] Alex Hanson, Allen Tu, Vasu Singla, Mayuka Jayawardhana, Matthias Zwicker, and Tom Goldstein. Pup 3dgs: Principled uncertainty pruning for 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 5949–5958, 2025.
- [25] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5664–5673, 2019.
- [26] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegrphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7515–7525, 2021.
- [27] Johanna Wald, Helisa Dharmo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3961–3970, 2020.
- [28] Hriday Bavle, Jose Luis Sanchez-Lopez, Muhammad Shaheer, Javier Civera, and Holger Voos. Situational graphs for robot navigation in structured indoor environments. *IEEE Robotics and Automation Letters*, 7(4):9107–9114, 2022.
- [29] Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, and Luca Carlone. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. In *Robotics: Science and Systems (RSS)*, 2020.
- [30] Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception engine for 3d scene graph construction and optimization. *Robotics: Science and Systems (RSS)*, 2022.
- [31] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [32] Krishan Rana, Jesse Haviland, Sourav Garg, Jad AbouChakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In *Conference on Robot Learning (CoRL)*, pages 23–72. PMLR, 2023.
- [33] Georgia Chalvatzaki, Ali Younes, Daljeet Nandha, An Thai Le, Leonardo FR Ribeiro, and Iryna Gurevych. Learning to reason over scene graphs: a case study of finetuning gpt-2 into a robot language model for grounded task planning. *Frontiers in Robotics and AI*, 10:1221739, 2023.
- [34] Zhe Ni, Xiaoxin Deng, Cong Tai, Xinyue Zhu, Qinghongbing Xie, Weihang Huang, Xiang Wu, and Long Zeng. Grid: Scene-graph-based instruction-driven robotic task planning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13765–13772, 2024.
- [35] Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems (NeurIPS)*, 37:5285–5307, 2024.
- [36] Lukas Schmid, Marcus Abate, Yun Chang, and Luca Carlone. Khronos: A unified approach for spatio-temporal metric-semantic slam in dynamic environments. 2024.
- [37] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 45–54, 2020.
- [38] Angela Dai and Matthias Nießner. Scan2mesh: From unstructured range scans to 3d meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5574–5583, 2019.
- [39] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 216–224, 2018.
- [40] Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *International conference on machine learning (ICML)*, pages 7220–7229. PMLR, 2020.
- [41] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 33:6840–6851, 2020.
- [42] Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19615–19625, 2024.
- [43] Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S Sukhatme. Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics*, 33(6):1273–1291, 2017.
- [44] Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.
- [45] Yao Zhao, Zhi Xiong, Shuailin Zhou, Jingqi Wang, Ling Zhang, and Pascual Campoy. Perception-aware planning for active slam in dynamic environments. *Remote Sensing*, 14(11):2584, 2022.
- [46] Mengya Xu, Yang Song, Yongbo Chen, Shoudong Huang, and Qi Hao. Invariant ekf based 2d active slam

- with exploration task. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5350–5356, 2021.
- [47] Ayoung Kim and Ryan M Eustice. Perception-driven navigation: Active visual slam for robotic area coverage. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3196–3203, 2013.
  - [48] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations (ICLR)*, 2020.
  - [49] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 42–48, 2024.
  - [50] Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, et al. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8120–8132, 2025.
  - [51] Junting Chen, Guohao Li, Suryansh Kumar, Bernard Ghanem, and Fisher Yu. How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers. 2024.
  - [52] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:4247–4258, 2020.
  - [53] Matthew Chang, Théophile Gervet, Mukul Khanna, Sri-ram Yenamandra, Dhruv Shah, So Yeon Min, Kavitha Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, et al. Goat: Go to any thing. In *Robotics: Science and Systems (RSS)*, 2024.
  - [54] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19595–19604, 2024.
  - [55] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, Andrew Y Ng, et al. Ros: an open-source robot operating system. In *IEEE International Conference on Robotics and Automation Workshop on Open Source Software*, volume 3, page 5, 2009.