COMPLEMENTARY GOAL GROUNDING VIA PROTOTYPE CONVOLUTION FOR OMNIGOAL NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

OmniGoal Navigation requires a single policy to locate goals in different forms, such as categories, images, or instance descriptions. Existing OmniGoal approaches primarily match heterogeneous goals with current observations, while navigation experience associated with different goal forms remains largely unmodeled. However, experience across goal forms can provide complementary cues for locating the same or related concepts, such as appearance from image goals and detailed textual descriptions from instance goals. To incorporate such complementary experience into a unified navigation policy, this work introduces Prototype Convolution (ProtoConv). A prototype graph organizes prior navigation experience into multi-form concept prototypes and their spatial relations. Given a goal, ProtoConv retrieves goal-conditioned prototype kernels and convolves them with an online semantic graph constructed from current observations. The convolution combines cross-form cues that transfer appearance and descriptive attributes for the same concept with cross-concept spatial correlations that support target localization through related observations. By integrating these cues, ProtoConv guides navigation with complementary experience across goal forms and concepts. Evaluations on HM3D, MP3D, and GOAT-Bench demonstrate consistent effectiveness across ObjectGoal, ImageGoal, and Instance Navigation without form-specific policy design.

1 INTRODUCTION

Goal-driven navigation requires an embodied agent to locate a specified target in an unknown environment (Anderson et al., 2018). Depending on the goal form, this problem is commonly studied as ObjectGoal Navigation (Zhang et al., 2024; 2025b), ImageGoal Navigation (Guo et al., 2025; Krantz et al., 2023), and Instance Navigation (Sun et al., 2024), where targets are specified through category names, reference images, and instance descriptions, respectively. Despite progress, many methods target individual goal forms (Yokoyama et al., 2024; Krantz et al., 2023), as illustrated in Figure 1(a). However, in practical applications, users may specify successive targets in different forms (Khanna et al., 2024). This motivates a single model for **OmniGoal Navigation** that handles heterogeneous goal specifications without switching among specialized policies.

Recent approaches have begun to explore this unified setting. One line of work encodes heterogeneous goals into shared implicit vision-language or structural representations, allowing a common navigation policy to operate across goal forms (Chang et al., 2024; Sun et al., 2024; Zhu et al., 2025). Another line builds maps or scene graphs that explicitly retain multi-form information for matching goals with observed scenes (Yin et al., 2025; Zhang et al., 2026; Huang et al., 2026). As summarized in Fig. 1(b), these approaches primarily address how different goals can be represented within a common navigation framework and matched with observations from the current environment. Such matching allows a unified policy to handle different goal forms, but the navigation experience associated with each form is largely used independently. However, experience acquired from different goal forms can provide complementary cues for locating the same or related concepts.

To this end, this work introduces **Prototype Convolution (ProtoConv)** for OmniGoal Navigation, as illustrated in Fig. 1(c). ProtoConv retrieves goal-conditioned prototype kernels from a prototype graph of prior navigation experience and convolves them with an online semantic graph constructed from current observations. Both structures preserve multi-form experience of individual concepts

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

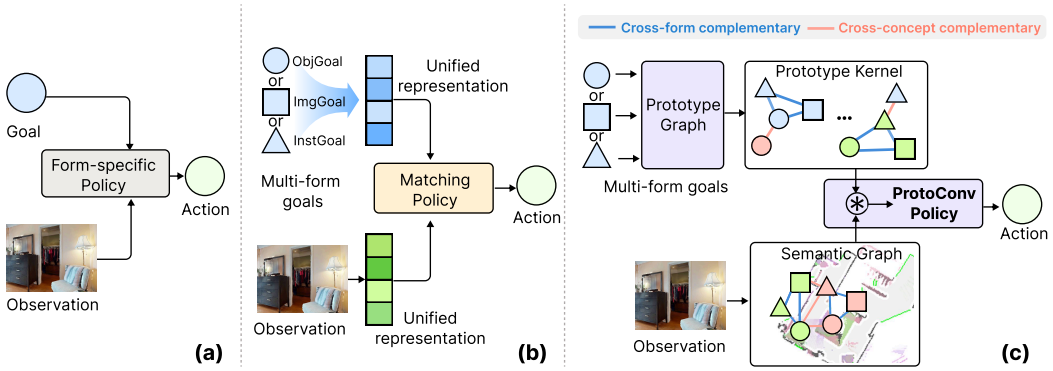

Figure 1: Comparison of goal-driven navigation policies. (a) Form-specific policy handles a single goal form. (b) Unified policy match heterogeneous goals to observations using implicit or explicit representations. (c) ProtoConv convolves goal-conditioned kernels retrieved from a prior prototype graph with the online semantic graph, exploiting cross-form and cross-concept complementarity.

and spatial relations among different concepts, enabling two complementary sources of navigation experience: **cross-form complementarity**, which transfers cues acquired from navigating to the same concept under different goal forms, e.g., fine-grained appearance from image goals or descriptive attributes from instance goals; and **cross-concept complementarity**, which exploits spatial correlations learned from related concepts, e.g., using an observed nightstand as contextual evidence for locating a bed. Therefore, navigation is guided by complementary experience across goal forms and concepts, rather than by goal–observation similarity alone.

The prototype graph aggregates category, text, and image prototypes, their cross-form associations, cross-concept relations, and observation counts from training environments. Semantic similarity and observation frequency select one root per form for each goal. Random walks with restart (Haveliwala, 2002) estimate root-centered relevance from node and relation counts, extracting connected prototype kernels that retain relevant context.

During navigation, observed instances and their multi-form content form an online semantic graph. Prototype convolution aligns each kernel root with online nodes and accumulates relevance-weighted, same-form affinities along relation-consistent paths, combining local evidence with associated forms and concepts. Kernel responses are projected onto the semantic map through Gaussian smoothing to guide waypoint selection. Evaluations on HM3D (Ramakrishnan et al., 2021; Yadav et al., 2023b), MP3D, and GOAT-Bench (Khanna et al., 2024) cover ObjectGoal, ImageGoal, and Instance Navigation. Across these tasks, ProtoConv matches or exceeds the compared specialized methods without form-specific policy design.

2 Related Work

Goal-driven Navigation. ObjectNav has developed from demonstration-based policies and semantic-map exploration (Ramrakhya et al., 2022; 2023; Chaplot et al., 2020; Ramakrishnan et al., 2022; Zhai & Wang, 2023) toward predictive planning and generalization across unseen objects and layouts (Zhang et al., 2024; Yu et al., 2024; Li et al., 2025; Zhang et al., 2022; 2023b). Open-vocabulary methods use pretrained vision–language representations and geometric cues (Radford et al., 2021; Gadre et al., 2023; Yokoyama et al., 2024; Yuan et al., 2024), together with commonsense semantics and memory-based reasoning (Zhou et al., 2023; Yu et al., 2023; Wu et al., 2024; Cao et al., 2025; Jin et al., 2026). Together, these approaches identify promising regions from category-level cues before the target is observed. Multimodal navigation extends goals to categories, images, and text descriptions. Shared representations and modular designs make these forms actionable within a common navigation framework (Majumdar et al., 2022; Sun et al., 2024; Krantz et al., 2023). Systems and benchmarks for heterogeneous goals also study persistent scene memory, graph matching, and frontier selection (Chang et al., 2024; Khanna et al., 2024; Yin et al., 2025; Zhu et al., 2025). Supporting several input forms leaves a separate question: whether evidence stored in one form can inform a goal specified in another. We focus on how a single-form goal can retrieve complementary evidence across forms and related concepts from accumulated cross-scene experience.

Structured Knowledge for Navigation. Structured navigation knowledge includes object relations, room layouts, and hierarchical object–zone–scene priors (Yang et al., 2019; Du et al., 2020; Wu et al., 2019; Zhang et al., 2021; 2025b). Other methods combine language and scene statistics or reason over online relations and functional associations (Chen et al., 2023; Yin et al., 2024; Zhang et al., 2025c). These representations encode correlations among objects, regions, and functions, providing search cues beyond target appearance. Multimodal scene memories range from grids and language-aligned 3D representations (Wang et al., 2023; 2025a; Wang & Lee, 2025) to descriptions, rendered views, and visual relations (Zhang et al., 2026; Huang et al., 2026). Multiview memories and graph-based records preserve retrievable evidence for navigation (Yang et al., 2025; Li et al., 2026; Niu et al., 2026). Their views and relations must still be selected for the current goal and compared with incomplete online observations. Retrieval-based approaches bring stored scene or trajectory experience into training and planning (Lewis et al., 2020; Wang et al., 2025b; 2026). ProtoConv instead organizes cross-scene image, text, category, and relational evidence into goal-conditioned response kernels that match local neighborhoods of the online graph.

3 Task Definition

Goal-driven navigation includes different task settings, such as ObjectGoal Navigation (Zhang et al., 2024), ImageGoal Navigation (Guo et al., 2025), and Instance Navigation (Sun et al., 2024). Most existing methods are designed for a specific task and handle only a single form of goal. In this work, we consider **OmniGoal Navigation**, following a similar setting to (Khanna et al., 2024; Yin et al., 2025), where the goal can be specified in different forms and a single model is expected to handle all of them.

At the beginning of each episode, an agent is placed in an unknown indoor environment without access to a complete map and is given one goal g of form $c_g \in \mathcal{C} = \{\text{cat}, \text{txt}, \text{img}\}$. At each step t, the agent receives an egocentric RGB observation I_t and predicts an action a_t. The discrete action space includes `move_forward`, `turn_left`, `turn_right`, and `stop`. The agent explores the environment online and autonomously issues `stop` to terminate navigation. An episode is considered successful if the agent stops within the maximum step at a valid goal location (e.g., less than $1m$), where the target is visible from the final viewpoint.

4 ProtoConv

Given a goal g specified in an arbitrary form (e.g., a category, instance description, or reference image), Our ProtoConv retrieves a set of goal-conditioned subgraphs from the prototype graph G as prototype kernels $\mathcal{K}_i$. Each kernel is centered on the queried goal while preserving its connections to spatially or semantically related prototypes, which provide complementary cues (e.g., spatial correlations or finer-grained details) to infer its likely location and accurately ground it in the environment.

At navigation step t, the agent incrementally maintains a semantic map $\mathcal{M}_t$ together with an online instance graph $\mathcal{G}_t$, which organizes observed objects in a structure analogous to the prototype graph. Each prototype kernel is convolved with $\mathcal{G}_t$, and the resulting responses are projected onto $\mathcal{M}_t$ to select the next waypoint, which is formulated as:

$$R_t = \Pi_{\mathcal{G}\to\mathcal{M}}\Big(\sum_i \mathcal{K}_i \circledast \mathcal{G}_t\Big), \qquad \mathcal{K}_i \sim p(\mathcal{K} \mid G, g), \tag{1}$$

where $\circledast$ denotes our prototype convolution, and $\Pi_{\mathcal{G}\to\mathcal{M}}$ projects graph-level responses to the spatial coordinates of $\mathcal{M}_t$. The distribution $p(\mathcal{K} \mid G, g)$ retrieves prototypes conditioned on the input goal g, the projected responses R_t indicate the spatial support of the retrieved prototypes on the semantic map. The next navigation waypoint is then selected from the location with the strongest response.

Note that our ProtoConv is not restricted to independently matching the queried goal. Both $\mathcal{K}_i$ and $\mathcal{G}_t$ contain heterogeneous goal representations, enabling complementary knowledge sharing along two dimensions. 1) **Cross-form complementarity**. Different forms of the same concept are associated to improve goal grounding. 2) **Cross-concept complementarity**. Spatial correlations are propagated across different concepts to infer likely goal locations. Consequently, even when g is

specified in only one form, the response can exploit complementary evidence from other forms and related concepts encoded in G.

4.1 PROTOTYPE GRAPH

To capture correlations across goal forms and concepts, we incrementally construct a prototype graph $G = (\mathcal{V}, \mathcal{E})$ from the training environments.

Graph Structure. Each node is defined as $v = (z_v, c_v, n_v) \in \mathcal{V}$, where z_v denotes the prototype content (e.g., category, instance description, or reference image), $c_v \in \{\text{cat}, \text{txt}, \text{img}\}$ specifies its goal form, and n_v records how frequently the prototype is observed. This node structure enables the same semantic concept to be represented by multiple form-specific prototypes.

Each edge is defined as $e(u, v) = (c_e, n_{u,v}) \in \mathcal{E}$, where u and v are two connected prototypes, c_e denotes their relation type, and $n_{u,v}$ records the frequency of this relation. Relation types include cross-form associations within the same concept and semantic or spatial relations across different concepts (e.g., "near" and "next to").

Incremental Construction. The prototype graph is incrementally constructed from RGB-D observations in the training environments. Given an new observation O_t, Grounding DINO (Liu et al., 2024) detects object instances and their categories, while MobileSAM (Zhang et al., 2023a) extracts instance masks. Combined with depth observations and camera intrinsics, the detected instances are projected onto the local semantic map m_t, which is progressively integrated into the global semantic map as $\mathcal{M}_t \leftarrow \mathcal{M}_t \cup m_t$. Based on the map integration results, newly observed instances are identified, and their corresponding masks are retained for subsequent relation graph construction.

A VLM predicts instance-level descriptions and semantic or spatial relations from the instance masks and corresponding images. The predictions are then organized into a local graph $\tilde{G}_t = (\tilde{\mathcal{V}}_t, \tilde{\mathcal{E}}_t)$, where each instance contributes category-, text-, and image-form nodes connected by cross-form edges, and different instances are connected according to their observed relations. The local graph is then incrementally merged into the global prototype graph by $G \leftarrow G \oplus \tilde{G}_t$, where $\oplus$ denotes prototype merging. Specifically, each local node is matched against existing prototypes of the same form. Matched nodes are merged and their occurrence counts are accumulated, while unmatched observations are inserted as new prototypes. During merging, textual prototypes are consolidated to preserve complementary instance details, while image prototypes retain the observation that best represents the corresponding concept. After node merging, local relations are mapped to the corresponding prototype pairs in G. Existing relations accumulate their observation counts, while unseen relations are added as new edges. By repeatedly merging local observations, G progressively aggregates recurring goal prototypes and their cross-form and cross-concept relations across training environments.

4.2 PROTOTYPE KERNEL

Given a goal g in an arbitrary form, a prototype kernel should capture not only its closest prototype, but also complementary cues from other forms and related concepts. Accordingly, three form-specific kernels are retrieved from G, corresponding to category, text, and image prototypes, respectively. Let $\mathcal{V}_c = \{v_i \in \mathcal{V} \mid c_v = c\}$ denote the prototypes of form $c \in \{\text{cat}, \text{txt}, \text{img}\}$. Since CLIP (Radford et al., 2021) maps visual and textual inputs into a shared embedding space, the input goal can directly query prototypes of all three forms, regardless of its original specification. For each form, the prototype root is selected by jointly considering semantic similarity and its observation frequency:

$$r_i = \arg\max_{v_i \in \mathcal{V}_c} \phi(g)^\top \phi(z_{v_i}) \log(1 + n_{v_i}) \tag{2}$$

where $\phi(\cdot)$ denotes the normalized CLIP embedding and n_v is the observation count. The count term gives more weight to prototypes that are observed more frequently.

Prototype Expansion. Given a goal-conditioned prototype r_i, the expansion aims to identify other prototypes that are relevant to the queried goal. This requires aggregating multi-path evidence around the root. Unlike fixed-hop or shortest-path expansion, random walks aggregate multi-path evidence and are employed to estimate a graph-wide relevance distribution. For an edge $e(u, v) = (c_e, n_{u,v})$, its reliability is measured by $a_{uv} = 2n_{uv}/(n_u + n_v)$, where n_{uv} is the edge

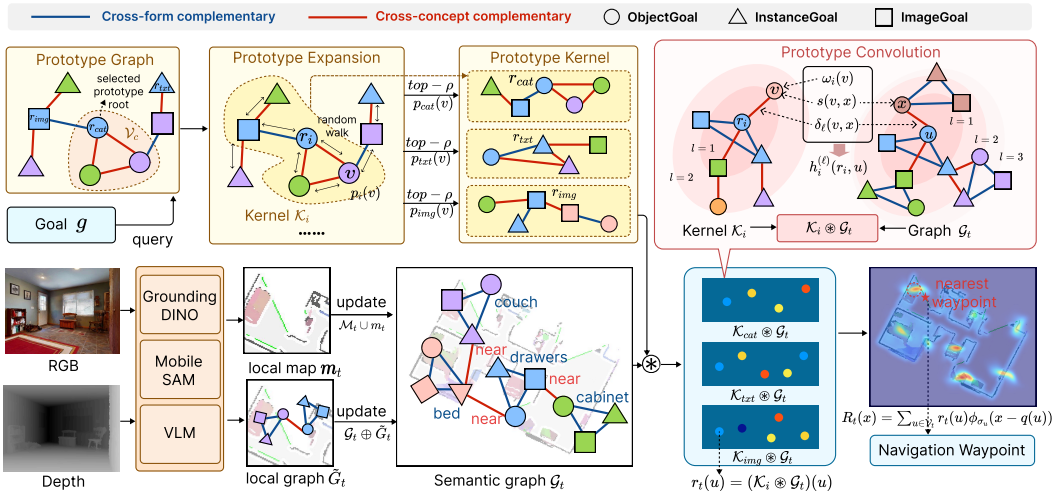

Figure 2: **Overview of ProtoConv.** Given a goal g in an arbitrary form, prototype roots are queried from the prototype graph and expanded to construct three prototype kernels. During navigation, each kernel is convolved with the online semantic graph $\mathcal{G}_t$ by matching same-form nodes along relation-consistent paths, integrating cross-form (blue) and cross-concept (red) complementary evidence. The resulting graph responses are aggregated and projected onto $\mathcal{M}_t$ for waypoint selection.

count and n_u, n_v are the occurrence counts of its endpoint prototypes. This normalization measures how consistently two prototypes co-occur while discounting their individual frequencies. The edge reliability is further normalized over $\mathcal{N}(u)$ to define the transition probability, where $\mathcal{N}(u)$ denotes the set of prototypes directly connected to u. The root-centered prototype distribution is then obtained by:

$$\mathbf{p}_i = \alpha \mathbf{e}_{r_i} + (1-\alpha)\mathbf{T}^\top \mathbf{p}_i, \qquad T_{uv} = \frac{\mathbb{I}[e(u,v)]\, a_{uv}}{\sum_{x \in \mathcal{N}(u)} a_{ux}} \tag{3}$$

where $\mathbf{e}_{r_i}$ is a one-hot vector with probability one at the root r_i and zero elsewhere, and α is the restart probability that prevents propagation from drifting too far from the root. Specifically, letting $\mathcal{V} = \{v_1, \ldots, v_{|\mathcal{V}|}\}$, the resulting distribution is $\mathbf{p}_i = [p_i(v_1), \ldots, p_i(v_{|\mathcal{V}|})]^\top$, where $p_i(v)$ measures the relevance of prototype v to the root r_i. The transition matrix is denoted as $\mathbf{T} = [T_{uv}]_{|\mathcal{V}| \times |\mathcal{V}|}$, where T_{uv} is the transition probability from prototype u to v. The distribution can be obtained by iteratively (Haveliwala, 2002) applying the above propagation from $\mathbf{p}_i^{(0)} = \mathbf{e}_{r_i}$ until convergence. This propagation is lightweight and performed only once at the beginning of each navigation episode. Consequently, $\mathbf{p}_i$ provides a probability-ranked support over the entire prototype graph, which is subsequently used to extract the root-centered prototype kernel.

Prototype Kernel. Given the root-centered distribution $\mathbf{p}_i$, the top-ρ fraction of prototypes are retained by probability, where $\rho \in (0, 1]$ controls the kernel expansion range. The resulting prototype kernel is obtained as

$$\mathcal{V}_i = \mathrm{CC}_{r_i}\left(\{v \in \mathcal{V} \mid p_i(v) \geq \rho\} \cup \{r_i\}\right), \qquad \mathcal{K}_i = G[\mathcal{V}_i] \tag{4}$$

Here, $\mathrm{CC}_{r_i}(\cdot)$ denotes the connected component containing r_i, and $G[\mathcal{V}_i]$ denotes the subgraph of G induced by $\mathcal{V}_i$. In this way, $\mathcal{K}_i$ retains the most relevant relational context around the root while excluding weakly associated prototypes.

4.3 Prototype Convolution

At the beginning of each navigation episode, given a goal g in an arbitrary form, three form-specific prototype kernels $\{\mathcal{K}_i\}_{i=1}^3$ are retrieved from the prototype graph G as described above. During navigation, an online semantic map $\mathcal{M}_t$ and a relation graph $\mathcal{G}_t$ are incrementally maintained from the agent observations. Given a new observation, the local semantic map m_t is integrated by $\mathcal{M}_t \leftarrow \mathcal{M}_t \cup m_t$, while the newly observed instances and their relations form a local graph $\tilde{G}_t$, which is

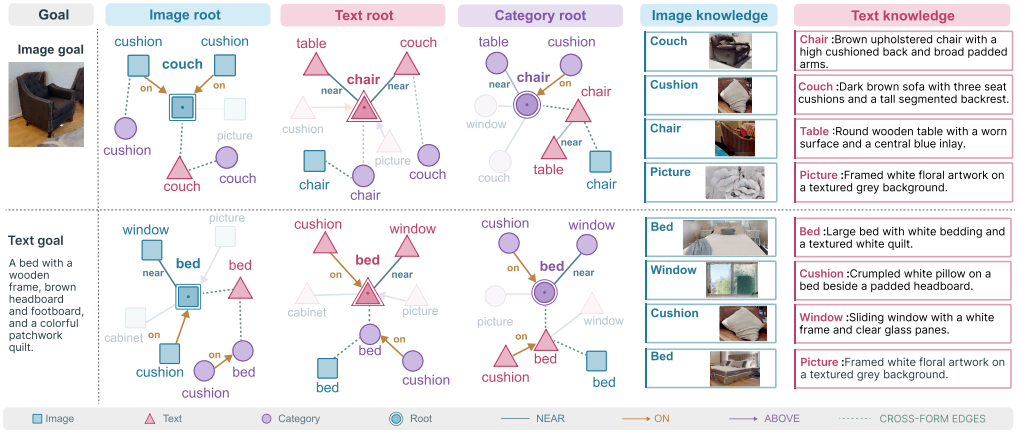

Figure 3: **Visualization of prototype kernels.** A single goal retrieves kernels containing complementary image, text, and category evidence. Double outlines mark roots; highlighted nodes and edges indicate retained knowledge, while faded elements are excluded.

merged into the online semantic graph as $\mathcal{G}_t \leftarrow \mathcal{G}_t \oplus \tilde{G}_t$, following the incremental construction in Sec. 4.1. The overall procedure is illustrated in Figure 2.

Convolution Computation. Given a prototype kernel $\mathcal{K}_i$ and online graph $\mathcal{G}_t$, prototype convolution slides the kernel root r_i over each online node u and measures their structural consistency. Each kernel node v is assigned a weight $\omega_i(v) = p_i(v) / \sum_{x \in \mathcal{V}_i} p_i(x)$, where $p_i(v)$ is obtained from prototype expansion. For a kernel node v and an online node x, their local affinity is defined as $s(v, x) = \mathbb{I}[c_v = c_x][\phi(z_v)^\top \phi(z_x)]_+$, where c_v denotes the node form and $[\cdot]_+ = \max(\cdot, 0)$.

For a sliding position u, let $\mathcal{V}_i^{(\ell)}$ denote the kernel nodes at relational depth ℓ from r_i, and $\mathcal{V}_t^{(\ell)}(u)$ denote the nodes at the same depth from u in $\mathcal{G}_t$. The structural response at depth ℓ is computed as

$$h_i^{(\ell)}(r_i, u) = \frac{1}{Z_\ell(u)} \sum_{v \in \mathcal{V}_i^{(\ell)}} \sum_{x \in \mathcal{V}_t^{(\ell)}(u)} \omega_i(v)\, \delta_\ell(v, x)\, s(v, x), \tag{5}$$

where $\delta_\ell(v, x)$ indicates whether the two edge paths have consistent relation types. Specifically, let $(\tau_1^v, \ldots, \tau_\ell^v)$ and $(\tau_1^x, \ldots, \tau_\ell^x)$ denote the relation types along the paths $r_i \to v$ and $u \to x$, respectively. Then $\delta_\ell(v, x) = \prod_{k=1}^{\ell} \mathbb{I}[\tau_k^v = \tau_k^x]$, such that only paths with identical relation sequences contribute to the response. $Z_\ell(u)$ is a normalization factor. Propagation continues until no further relation-consistent nodes are available in either $\mathcal{K}_i$ or $\mathcal{G}_t$, yielding a set of valid relational depths denoted by $\mathcal{D}_i(u)$. The prototype convolution response at u is then defined as

$$r_t(u) = (\mathcal{K}_i \circledast \mathcal{G}_t)(u) = \frac{1}{|\mathcal{D}_i(u)|} \sum_{\ell \in \mathcal{D}_i(u)} h_i^{(\ell)}(r_i, u), \qquad u \in \mathcal{V}_t. \tag{6}$$

The resulting responses over $\mathcal{G}_t$ are projected onto the semantic map for waypoint selection.

Navigation Waypoint. The responses from all prototype kernels are first aggregated over the online graph as $\sum_i (\mathcal{K}_i \circledast \mathcal{G}_t)$ and then projected onto the semantic map by $R_t = \Pi_{\mathcal{G} \to \mathcal{M}}(\sum_i \mathcal{K}_i \circledast \mathcal{G}_t)$. Specifically, the projected response at location x is computed by

$$R_t(x) = \sum_{u \in \mathcal{V}_t} r_t(u)\, \phi_{\sigma_u}(x - q(u)), \tag{7}$$

where $r_t(u)$ denotes the aggregated response at node u, $q(u)$ is its coordinate on the semantic map, and $\phi_\sigma(d) = \exp(-\|d\|_2^2 / 2\sigma^2)$ denotes a Gaussian kernel. The diffusion scale σ_u is determined by the spatial extent of the incident edges, adapting the projection to the local graph structure without a fixed smoothing radius. The highest-response location is then obtained by $\hat{w}_t = \arg\max_{x \in \mathcal{M}_t} R_t(x)$. At the early stage of navigation, observations are sparse and the frontier closest to $\hat{w}_t$ is selected to encourage exploration. After this warm-up stage, the response maximum is used as the navigation waypoint.

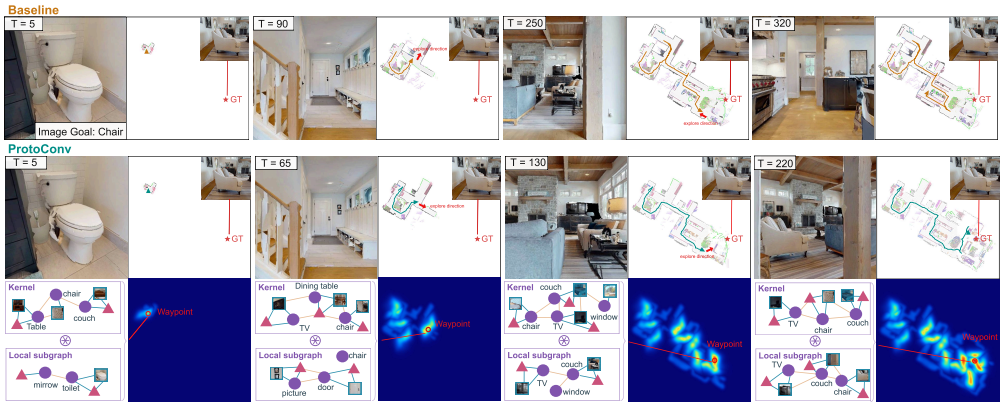

Figure 4: **ImageNav with ProtoConv.** Prototype kernels are convolved with local subgraphs of the online semantic graph, integrating cross-form and cross-concept cues to guide waypoint selection.

5 Experiments

5.1 Experimental Setup

Datasets and Metrics. We evaluate ProtoConv in Habitat (Savva et al., 2019) on HM3D (Ramakrishnan et al., 2021; Yadav et al., 2023b) and MP3D. ObjectNav uses HM3Dv1, HM3Dv2, and MP3D validation benchmarks, each with six target categories; image-goal navigation (ImageNav) and text-goal Instance Navigation (InsNav) use HM3D. GOAT-Bench Val Unseen (Khanna et al., 2024) tests sequential category, image, and text goals. We report success rate (SR) and success weighted by path length (SPL) (Anderson et al., 2018). For GOAT-Bench, each subgoal is scored separately, with path length measured from its starting position.

Implementation Details. Grounding DINO (Liu et al., 2024) detects object instances and their categories, and MobileSAM (Zhang et al., 2023a) extracts instance masks. We use CLIP (Radford et al., 2021) to encode image and text inputs, and Qwen3.7-Plus to generate instance descriptions and identify semantic and spatial relations. The restart probability is $\alpha = 0.5$ and the kernel truncation threshold is $\rho = 0.6$. HM3D episodes allow up to 500 actions, with 0.25 m forward steps and 30° turns. Success is evaluated at the stop action against valid goal viewpoints within 0.2 m.

5.2 Evaluation Results

Visualization of ProtoConv. Figure 3 shows how a single-form goal retrieves knowledge across forms and relations. For the chair image goal, the image root is a couch, but the text and category roots are chairs; the kernels also retain relations such as a cushion ON a couch. Cross-form roots and object relations therefore supply evidence beyond the goal's input form.

Figure 4 compares the navigation trajectories of ProtoConv and the baseline. In frame 2, the baseline enters a non-target region, while weak kernel matching assigns that branch a low response and ProtoConv selects another waypoint. In frame 3, image and text evidence and relations among chairs, sofas, and televisions produce a higher response in the living area; ProtoConv moves toward the target chair while the baseline backtracks. The lower panels show responses from matching the same kernel to successive local subgraphs, using cross-form evidence and object relations to reach the target region sooner and avoid unproductive exploration.

Ablation of Knowledge Forms. Table 1 tests whether supplementing knowledge in different forms benefits goals specified in each form. Successively adding category, image, and text knowledge (IDs 2–5) improves both SR and SPL across ObjectNav, ImageNav, and InsNav. Image knowledge improves text-goal navigation, and text knowledge improves image-goal navigation. These cross-form gains directly validate our central motivation: knowledge acquired in different forms provides complementary evidence for goal grounding.



Figure 5: **Joint sensitivity to α and ρ on HM3D Instance Navigation.** Rows vary the restart probability α and columns vary the kernel truncation parameter ρ. Cells report SR (left) and SPL (right) in percent; darker cells indicate higher values.

IDs 1 and 5 further compare VLM-supplied knowledge with our Prototype Graph while retaining the same three forms. The VLM supplies general associations; our graph preserves finer-grained, observation-grounded prototypes, their cross-form and cross-concept relations, and relation-support frequencies. Its consistent advantage across all three tasks demonstrates that this finer-grained organization makes complementary knowledge more effective than general VLM supplementation.

Table 1: Knowledge-form ablation (%). C/I/T: category/image/text form. ObjNav: ObjectGoal Navigation (HM3Dv2); ImgNav: ImageGoal Navigation (HM3D); InsNav: Instance Navigation with text goals (HM3D).

ID	Source	C	I	T	ObjNav		ImgNav		InsNav	
					SR	SPL	SR	SPL	SR	SPL
1	VLM	✓	✓	✓	76.7	37.3	61.5	29.8	30.2	11.2
2	Prototype Graph	×	×	×	70.3	32.3	59.6	28.1	26.1	10.6
3		✓	×	×	72.6	36.9	60.4	28.9	27.7	11.5
4		✓	✓	×	76.9	38.4	62.5	32.1	29.7	11.7
5		✓	✓	✓	**78.5**	**40.5**	**65.7**	**33.7**	**31.4**	**13.3**

Ablation of Kernel Construction and Response. Table 2 separates how complementary knowledge is selected from how it is matched. Under both response functions, random-walk kernels outperform fixed two-hop kernels across all tasks. Following historically supported relations gathers other forms of the goal and related concepts beyond a fixed neighborhood. These results establish the advantage of selecting complementary knowledge through relational support rather than a prescribed expansion depth.

Table 2: Kernel-construction and response ablation (%). Fixed uses a probability-ranked two-hop neighborhood; Prototype Kernel uses random-walk expansion. Node matches node content; Node+Edge also matches relation types along paths.

ID		Kernel	Response	ObjNav		ImgNav		InsNav	
				SR	SPL	SR	SPL	SR	SPL
I	1	Fixed	Node	74.3	35.1	59.6	28.9	27.6	10.2
	2		Node+Edge	77.1	38.9	64.7	32.9	30.7	12.2
II	3	Prototype Kernel	Node	76.9	37.8	61.4	31.7	28.9	11.9
	4		Node+Edge	**78.5**	**40.5**	**65.7**	**33.7**	**31.4**	**13.3**

The Node versus Node+Edge comparisons (IDs 1–2 and 3–4) isolate the relation-consistency term $\delta_\ell(v, x)$ in Eq. (5). This term admits node affinity only when the kernel and online paths have matching edge-relation types at every step. Its gains under both kernel constructions demonstrate that node similarity alone is insufficient: preserving relational context improves goal grounding, especially for image- and text-specified instances.

Hyperparameters tuning. Figure 5 evaluates how two kernel-construction hyperparameters jointly affect SR and SPL on HM3D text-goal Instance Navigation. The restart probability α controls how often the walk returns to the prototype root; larger values concentrate relevance near that root. The truncation threshold ρ controls node retention, with larger values producing smaller kernels. Both metrics peak at $\alpha = 0.5$ and $\rho = 0.6$, supporting a balance between retaining complementary context and excluding weakly relevant knowledge.

5.3 Comparison with SOTA Methods

Object-goal Navigation. A category goal specifies what to find but leaves its appearance and surrounding context unspecified. Table 3 shows that ProtoConv leads the listed methods in SR and SPL across HM3Dv1, HM3Dv2, and MP3D. Together with the knowledge-form ablation, these results

Table 3: Comparison with related methods on ObjectNav in HM3Dv1, HM3Dv2, and MP3D (%). OV denotes support for goals beyond a fixed category vocabulary; MG denotes support for multiple goal forms.

Method	OV	MG	HM3Dv1		HM3Dv2		MP3D	
			SR	SPL	SR	SPL	SR	SPL
DD-PPO (Wijmans et al., 2020)	×	×	27.9	14.2	-	-	8.0	1.8
SGM (Zhang et al., 2024)	×	×	60.2	30.8	-	-	37.7	14.7
ESC (Zhou et al., 2023)	✓	×	39.2	22.3	-	-	28.7	14.2
VLFM (Yokoyama et al., 2024)	✓	×	52.5	30.4	63.6	32.5	36.4	17.5
VoroNav (Wu et al., 2024)	✓	×	42.0	26.0	-	-	-	-
SG-Nav (Yin et al., 2024)	✓	×	54.0	24.9	49.6	25.5	40.2	16.0
ApexNav (Zhang et al., 2025a)	✓	×	59.6	33.0	76.2	38.0	39.2	17.8
BeliefMapNav (Zhou et al., 2025)	✓	×	61.4	30.6	-	-	37.3	17.6
TrajRAG (Wang et al., 2026)	✓	×	62.5	33.9	78.1	40.2	42.6	18.0
ZSON (Majumdar et al., 2022)	✓	✓	25.5	12.6	-	-	15.3	4.8
PSL (Sun et al., 2024)	✓	✓	42.4	19.2	-	-	-	-
InstructNav (Long et al., 2025)	✓	✓	-	-	58.0	20.9	-	-
UniGoal (Yin et al., 2025)	✓	✓	54.5	25.1	-	-	41.0	16.4
MSGNav (Huang et al., 2026)	✓	✓	-	-	74.1	33.4	-	-
ProtoConv (Ours)	✓	✓	**63.5**	**34.2**	**78.5**	**40.5**	**43.0**	**18.7**

demonstrate that image and text knowledge benefits navigation even when the goal is given only as a category. A single-form goal therefore need not restrict the forms of knowledge used to ground it.

Instance Navigation. Locating a specific instance requires finer discrimination than identifying its category. ProtoConv achieves the highest SR and SPL for both image- and text-specified goals in Table 4. Together with the form and edge-consistency ablations, these results demonstrate the value of combining complementary appearance, descriptions, and relational context for instance grounding. The same prototype graph supports this finer discrimination across both goal forms.

Table 4: Comparison with related methods on image- and text-specified instance navigation in HM3D (%).

Method	OV	MG	ImgNav		InsNav	
			SR	SPL	SR	SPL
IIN-RL (Krantz et al., 2022)	×	×	8.3	3.5	–	–
OVRL-v2-IIN (Yadav et al., 2023a)	×	×	24.8	11.8	–	–
Mod-IIN (Krantz et al., 2023)	✓	×	56.1	23.3	–	–
PSL (Sun et al., 2024)	✓	✓	23.0	11.4	16.5	7.5
GOAT (RL) (Chang et al., 2024)	✓	✓	37.4	16.1	17.0	8.8
UniGoal (Yin et al., 2025)	✓	✓	60.2	23.7	20.2	11.4
ProtoConv (Ours)	✓	✓	**65.7**	**33.7**	**31.4**	**13.3**

Sequential-goal Navigation. GOAT-Bench tests whether shared knowledge remains useful as successive goals change between category, image, and text forms. The leading SR and SPL in Table 5 demonstrate that ProtoConv handles these changes within a common navigation framework. The benefit of complementary prototype knowledge thus extends beyond isolated goals to continued navigation across goal forms, supporting our motivation for a shared knowledge representation.

Table 5: Comparison with related methods on GOAT-Bench Val Unseen (%).

Method	MG	Overall	
		SR	SPL
SenseAct-NN Monolithic (Khanna et al., 2024)	✓	12.3	6.8
Modular CLIP on Wheels (Gadre et al., 2023)	✓	16.1	10.4
VLMnav (Goetting et al., 2025)	✓	20.1	9.6
GOAT (Modular) (Chang et al., 2024)	✓	24.9	17.2
DyNaVLM (Ji et al., 2025)	✓	25.5	10.2
3D-Mem (Yang et al., 2025)	✓	28.8	15.8
SenseAct-NN Skill Chain (Khanna et al., 2024)	✓	29.5	11.3
TANGO (Ziliotto et al., 2025)	✓	32.1	16.5
MTU3D (Zhu et al., 2025)	✓	47.2	27.7
MSGNav (Huang et al., 2026)	✓	52.0	29.6
ProtoConv (Ours)	✓	**53.1**	**30.7**

6 Conclusion

We presented ProtoConv for OmniGoal Navigation, integrating prior navigation experience across goal forms and concepts into a unified policy. Goal-conditioned prototype kernels retrieved from a prototype graph are convolved with the online semantic graph, combining cross-form cues and cross-concept spatial correlations. Experiments on HM3D, MP3D, and GOAT-Bench demonstrate effective navigation across goal forms without form-specific policies. These results highlight complementary experience as evidence for goal grounding beyond goal–observation similarity.

AI USE STATEMENT

We used generative AI tools to assist with English translation and language polishing, LaTeX formatting and presentation, and brainstorming and discussion during manuscript preparation. The research methodology, experimental results, and final technical claims were determined and verified by the authors. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final manuscript.

REFERENCES

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On Evaluation of Embodied Navigation Agents. *arXiv preprint arXiv:1807.06757*, 2018. doi: 10.48550/arXiv.1807.06757.

Yihan Cao, Jiazhao Zhang, Zhinan Yu, Shuzhen Liu, Zheng Qin, Qin Zou, Bo Du, and Kai Xu. CogNav: Cognitive Process Modeling for Object Goal Navigation with LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9550–9560, 2025.

Matthew Chang, Theophile Gervet, Mukul Khanna, Sriram Yenamandra, Dhruv Shah, So Yeon Min, Kavit Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, Roozbeh Mottaghi, Jitendra Malik, and Devendra Singh Chaplot. GOAT: GO to Any Thing. In *Robotics: Science and Systems XX*, 2024. doi: 10.15607/RSS.2024.XX.073.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object Goal Navigation using Goal-Oriented Semantic Exploration. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4247–4258, 2020.

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. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. doi: 10.15607/RSS.2023.XIX.075.

Heming Du, Xin Yu, and Liang Zheng. Learning Object Relation Graph and Tentative Policy for Visual Navigation. In *Computer Vision – ECCV 2020*, volume 12352 of *Lecture Notes in Computer Science*, pp. 19–34. Springer, 2020. doi: 10.1007/978-3-030-58571-6_2.

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)*, pp. 23171–23181, 2023.

Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. End-to-End Navigation with Vision-Language Models: Transforming Spatial Reasoning into Question-Answering. In *Proceedings of the International Conference on Neuro-symbolic Systems*, volume 288 of *Proceedings of Machine Learning Research*, pp. 22–35. PMLR, 2025.

Wenxuan Guo, Xiuwei Xu, Hang Yin, Ziwei Wang, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Igl-nav: incremental 3d gaussian localization for image-goal navigation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6808–6817. IEEE, 2025.

Taher H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the 11th International World Wide Web Conference*, 2002.

Xun Huang, Shijia Zhao, Yunxiang Wang, Xin Lu, Wanfa Zhang, Rongsheng Qu, Weixin Li, Yunhong Wang, and Chenglu Wen. MSGNav: Unleashing the Power of Multi-modal 3D Scene Graph for Zero-Shot Embodied Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37154–37163, 2026.

Zihe Ji, Huangxuan Lin, and Yue Gao. DyNaVLM: Zero-Shot Vision-Language Navigation System with Dynamic Viewpoints and Self-Refining Graph Memory. *arXiv preprint arXiv:2506.15096*, 2025. doi: 10.48550/arXiv.2506.15096.

Qunchao Jin, Yilin Wu, and Changhao Chen. PanoNav: Mapless Zero-Shot Object Navigation with Panoramic Scene Parsing and Dynamic Memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(22):18351–18359, 2026. doi: 10.1609/aaai.v40i22.38899.

Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. GOAT-Bench: A Benchmark for Multi-Modal Lifelong Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16373–16383, 2024.

Jacob Krantz, Stefan Lee, Jitendra Malik, Dhruv Batra, and Devendra Singh Chaplot. Instance-Specific Image Goal Navigation: Training Embodied Agents to Find Object Instances. *arXiv preprint arXiv:2211.15876*, 2022. doi: 10.48550/arXiv.2211.15876.

Jacob Krantz, Theophile Gervet, Karmesh Yadav, Austin Wang, Chris Paxton, Roozbeh Mottaghi, Dhruv Batra, Jitendra Malik, Stefan Lee, and Devendra Singh Chaplot. Navigating to Objects Specified by Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10916–10925, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Badi Li, Ren-Jie Lu, Yu Zhou, Jingke Meng, and Wei-Shi Zheng. Distilling LLM Prior to Flow Model for Generalizable Agent's Imagination in Object Goal Navigation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Mingyi Li, Hui Liu, Ying Li, Shubo Zhang, Chunle Gao, Xiaokang Ma, Hanqing Hu, and Weixin Mao. EMKG: Embodied Memory Knowledge Graphs for Object-Goal Navigation in Dynamic Open Worlds. *IEEE Robotics and Automation Letters*, 11(4):4537–4544, 2026. doi: 10.1109/LRA.2026.3655297.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024.

Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. InstructNav: Zero-shot System for Generic Instruction Navigation in Unexplored Environment. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2049–2060. PMLR, 2025.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. ZSON: Zero-Shot Object-Goal Navigation using Multimodal Goal Embeddings. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32340–32352. Curran Associates, Inc., 2022. doi: 10.52202/068431-2343.

Haochen Niu, Lantao Zhang, Xingwu Ji, Rendong Ying, Peilin Liu, and Fei Wen. SSMG-Nav: Enhancing Lifelong Object Navigation with Semantic Skeleton Memory Graph. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. *arXiv preprint arXiv:2109.08238*, 2021. doi: 10.48550/arXiv.2109.08238.

Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. PONI: Potential Functions for ObjectGoal Navigation with Interaction-free Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18890–18900, 2022.

Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-Web: Learning Embodied Object-Search Strategies From Human Demonstrations at Scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5173–5183, 2022.

Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. PIRLNav: Pretraining with Imitation and RL Finetuning for ObjectNav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17896–17906, 2023.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9339–9347, 2019.

Xinyu Sun, Lizhao Liu, Hongyan Zhi, Ronghe Qiu, and Junwei Liang. Prioritized Semantic Learning for Zero-shot Instance Navigation. In *Computer Vision – ECCV 2024*, volume 15070 of *Lecture Notes in Computer Science*, pp. 161–178. Springer, 2024. doi: 10.1007/978-3-031-73254-6_10.

Yiyao Wang, Sixian Zhang, Keming Zhang, Xinhang Song, Songjie Du, and Shuqiang Jiang. TrajRAG: Retrieving Geometric-Semantic Experience for Zero-Shot Object Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15166–15176, 2026.

Zihan Wang and Gim Hee Lee. g3D-LF: Generalizable 3D-Language Feature Fields for Embodied Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14191–14202, 2025.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. GridMM: Grid Memory Map for Vision-and-Language Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15625–15636, 2023.

Zihan Wang, Seungjun Lee, and Gim Hee Lee. Dynam3D: Dynamic Layered 3D Tokens Empower VLM for Vision-and-Language Navigation. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Zihan Wang, Yaohui Zhu, Gim Hee Lee, and Yachun Fan. NavRAG: Generating User Demand Instructions for Embodied Navigation through Retrieval-Augmented LLM. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8430–8440. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.findings-acl.442.

Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames. In *International Conference on Learning Representations (ICLR)*, 2020.

Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. VoroNav: Voronoi-based Zero-shot Object Navigation with Large Language Model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53757–53775. PMLR, 2024.

Yi Wu, Yuxin Wu, Aviv Tamar, Stuart Russell, Georgia Gkioxari, and Yuandong Tian. Bayesian Relational Memory for Semantic Visual Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2769–2779, 2019. doi: 10.1109/ICCV.2019.00286.

Karmesh Yadav, Arjun Majumdar, Ram Ramrakhya, Naoki Yokoyama, Alexei Baevski, Zsolt Kira, Oleksandr Maksymets, and Dhruv Batra. OVRL-V2: A simple state-of-art baseline for ImageNav and ObjectNav. *arXiv preprint arXiv:2303.07798*, 2023a. doi: 10.48550/arXiv.2303.07798.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-Matterport 3D Semantics Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4927–4936, 2023b.

Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual Semantic Navigation using Scene Priors. In *International Conference on Learning Representations (ICLR)*, 2019.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3D-Mem: 3D Scene Memory for Embodied Exploration and Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17294–17303, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 5285–5307. Curran Associates, Inc., 2024. doi: 10.52202/079017-0171.

Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. UniGoal: Towards Universal Zero-shot Goal-oriented Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19057–19066, 2025.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 42–48, 2024. doi: 10.1109/icra57147.2024.10610712.

Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3MVN: Leveraging Large Language Models for Visual Target Navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3554–3560, 2023. doi: 10.1109/IROS55552.2023.10342512.

Xinyao Yu, Sixian Zhang, Xinhang Song, Xiaorong Qin, and Shuqiang Jiang. Trajectory Diffusion for ObjectGoal Navigation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110388–110411, 2024. doi: 10.52202/079017-3504.

Shuaihang Yuan, Hao Huang, Yu Hao, Congcong Wen, Anthony Tzes, and Yi Fang. GAMap: Zero-Shot Object Goal Navigation with Multi-Scale Geometric-Affordance Guidance. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39386–39408, 2024. doi: 10.52202/079017-1244.

Albert J. Zhai and Shenlong Wang. PEANUT: Predicting and Navigating to Unseen Targets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10926–10935, 2023.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023a.

Mingjie Zhang, Yuheng Du, Chengkai Wu, Jinni Zhou, Zhenchao Qi, Jun Ma, and Boyu Zhou. ApexNav: An Adaptive Exploration Strategy for Zero-Shot Object Navigation with Target-centric Semantic Fusion. *IEEE Robotics and Automation Letters*, 10(11):11530–11537, 2025a. doi: 10.1109/LRA.2025.3606388.

Sixian Zhang, Xinhang Song, Yubing Bai, Weijie Li, Yakui Chu, and Shuqiang Jiang. Hierarchical Object-to-Zone Graph for Object Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15130–15140, 2021.

Sixian Zhang, Weijie Li, Xinhang Song, Yubing Bai, and Shuqiang Jiang. Generative Meta-Adversarial Network for Unseen Object Navigation. In *Computer Vision – ECCV 2022*, pp. 301–320. Springer, 2022. doi: 10.1007/978-3-031-19842-7_18.

Sixian Zhang, Xinhang Song, Weijie Li, Yubing Bai, Xinyao Yu, and Shuqiang Jiang. Layout-based Causal Inference for Object Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10792–10802, 2023b. doi: 10.1109/CVPR52729.2023.01039.

Sixian Zhang, Xinyao Yu, Xinhang Song, Xiaohan Wang, and Shuqiang Jiang. Imagine Before Go: Self-Supervised Generative Map for Object Goal Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16414–16425, 2024.

Sixian Zhang, Xinhang Song, Xinyao Yu, Yubing Bai, Xinlong Guo, Weijie Li, and Shuqiang Jiang. HOZ++: Versatile Hierarchical Object-to-Zone Graph for Object Navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5958–5975, 2025b. doi: 10.1109/TPAMI.2025.3552987.

Sixian Zhang, Xinyao Yu, Xinhang Song, Yiyao Wang, and Shuqiang Jiang. Function-centric Bayesian Network for Zero-Shot Object Goal Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19535–19545, 2025c.

Sixian Zhang, Yiyao Wang, Xinhang Song, Keming Zhang, Zijian Xu, and Shuqiang Jiang. Multi-Scale Gaussian-Language Map for Zero-shot Embodied Navigation and Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37086–37097, 2026.

Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. ESC: Exploration with Soft Commonsense Constraints for Zero-shot Object Navigation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42829–42842. PMLR, 2023.

Zibo Zhou, Yue Hu, Lingkai Zhang, Zonglin Li, and Siheng Chen. BeliefMapNav: 3D Voxel-Based Belief Map for Zero-Shot Object Navigation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 83008–83036. Curran Associates, Inc., 2025. doi: 10.52202/085713-2782.

Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, Siyuan Huang, and Qing Li. 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)*, pp. 8120–8132, 2025.

Filippo Ziliotto, Tommaso Campari, Luciano Serafini, and Lamberto Ballan. TANGO: Training-free Embodied AI Agents for Open-world Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24603–24613, 2025.

A Appendix

This appendix expands the graph construction, root-centered expansion, and local matching described in the main paper.

A.1 Observation graphs and incremental aggregation

For a tracked object instance o, the local graph contains the available form-specific nodes o^c, o^t, and o^i for category, description, and image evidence. Same-instance edges connect these nodes. Such edges express common instance identity and are distinct from physical relations between two objects. The image node retains an observed crop and its feature; the text node retains the description and its feature; the category node retains the category string and its text feature. Missing or unusable evidence does not require a fabricated form node.

For two tracked instances o and o', a physical relation is admitted only when its endpoints and type can be grounded in the observation. The local graph may link nodes of the same or different forms across the two instances; the relation label retains its meaning under either representation. We

use the finite physical-relation vocabulary NEAR, ON, ABOVE, INSIDE, and ATTACHED-TO when supported by the available annotation pipeline. A functional-area tag may help interpret the instance description or prototype identity, but does not become an area node or an area-membership edge.

Let $m(o^f)$ map an observed node of form f to an existing same-form prototype or to a newly inserted prototype. The occurrence count is updated as

$$n_{m(o^f)} \leftarrow n_{m(o^f)} + 1. \tag{8}$$

Each grounded local edge $(o^f, o'^{f'}, r)$ is mapped in the same update to $(m(o^f), m(o'^{f'}), r)$; its support count is then incremented or the edge is inserted. Thus node consolidation and relation accumulation share the same instance-to-prototype mapping. When one endpoint is new, its relation is retained rather than discarded. Repeated views of one tracked physical pair should be deduplicated under a stable evidence identifier before an edge count is interpreted as independent support. This counting convention requires the final export to report its deduplication key.

A.2 Retrieval and restart-walk derivation

For a goal g, each form $f \in \{c, t, i\}$ selects one root among that form's prototypes using the root-selection score defined in the main paper. CLIP's shared image–text space permits direct comparison with all three form collections; the retrieval does not require generating a synthetic goal in either missing form. The logarithmic count factor favors recurrent prototypes without replacing content similarity.

For a nonzero edge count n_{uv}, write $q_{u\to v} = n_{uv}/n_u$ and $q_{v\to u} = n_{uv}/n_v$. The reliability used in the main text is their harmonic mean:

$$\frac{2}{q_{u\to v}^{-1} + q_{v\to u}^{-1}} = \frac{2n_{uv}}{n_u + n_v} = a_{uv}. \tag{9}$$

This makes an edge less influential if it is common only because one endpoint is very frequent. The row-normalized transition matrix T satisfies $\sum_v T_{uv} = 1$ for nonisolated nodes. An isolated root can be assigned a self-loop so that its row is also stochastic.

Starting at root r, the restart-walk fixed point is $\mathbf{p} = \alpha\mathbf{e}_r + (1-\alpha)T^\top\mathbf{p}$. Since $0 < \alpha \le 1$ and T is stochastic, it has the convergent expansion

$$\mathbf{p} = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k (T^\top)^k \mathbf{e}_r. \tag{10}$$

The kth term collects all length-k paths from the root, weighted by their transition probabilities and the restart factor. This explains why several moderate supporting paths can reinforce a prototype without requiring a single shortest path. The weights sum to one: $\mathbf{1}^\top\mathbf{p} = \alpha + (1-\alpha)\mathbf{1}^\top\mathbf{p}$, hence $\mathbf{1}^\top\mathbf{p} = 1$.

Given a probability cutoff ρ, define $S_{r,\rho} = \{v : p_r(v) \ge \rho\} \cup \{r\}$. The kernel is the subgraph induced by the root-containing connected component of $S_{r,\rho}$. Consequently every retained node has a path to the root entirely inside the kernel. The cutoff controls the admitted support; it is not a fixed hop count or fixed number of neighbors. The retained component may overlap with a kernel rooted at another form, while the roots and the admitted evidence can differ.

A.3 Node and relation responses

At an online graph node u, the kernel root is aligned with u. A kernel node v and online node x first require matching forms and then contribute their nonnegative embedding affinity $s(v, x)$. For a depth-ℓ path pair, the relation indicator is

$$\delta_\ell(v, x) = \prod_{j=1}^{\ell} \mathbb{I}[\tau_j^v = \tau_j^x]. \tag{11}$$

At depth zero, the empty product equals one, so the root response reduces to node affinity. At positive depth, a mismatch at any relation step suppresses the path-pair contribution. The node-only

ablation uses the same kernel nodes but sets $\delta_\ell \equiv 1$; this isolates relation consistency at matching time from the effect of relation-weighted kernel construction.

The main paper's depth response weights a matched kernel node by its normalized walk probability $\omega_r(v) = p_r(v)/\sum_{z\in\mathcal{V}_r} p_r(z)$. Averaging valid depth responses gives the graph response at u. The map response is the superposition of Gaussian contributions at the observed node coordinates:

$$R_t(x) = \sum_{u\in\mathcal{V}_t} r_t(u)\exp\left(-\frac{\|x-q(u)\|_2^2}{2\sigma_u^2}\right). \tag{12}$$

The spatial scale σ_u is tied to the extent of incident observed relations. The active frontier is chosen relative to the response maximum during warm-up, after which the maximum itself is used as the waypoint, as specified in the main text.

B Graph-Scale Projection and Evidence Accounting

The full-corpus evidence inventory contains 66,288 object-evidence records. The resulting Prototype Graph contains 1557 prototype nodes and 4,748 edges. Each retained prototype is represented by the available category, image, and text forms. Distinct modes within the same category may be preserved as separate prototypes when their evidence exhibits clear differences, allowing the graph to retain intra-category diversity rather than collapsing all observations into a single prototype.

C Vision–Language Annotation Protocol

The VLM supplies descriptions and candidate relations from observed images; it does not provide ground-truth graph edges. The templates below summarize the instructions found in the construction scripts. Bracketed fields are runtime inputs, not literal text. Outputs are parsed and checked for known instance aliases, allowed relation types, and explicit uncertainty. The final submission should attach the exact frozen prompt version, model name, and decoding configuration used to build its reported graph.

C.1 Instance description

For an object crop, the description prompt asks the model to describe only intrinsic visible appearance and return structured attributes:

```
Describe only the visible object intrinsic appearance.
Category hint: [CATEGORY]. Return JSON with description,
attributes, materials, and colors. Do not infer neighbors
or rooms.
```

This separation matters because a crop may identify the object but cannot establish its physical relation to another object outside the crop.

C.2 Grounded inter-instance relations

The relation prompt receives a full RGB frame with indexed boxes $O0, O1, \ldots$, optional crops, and the object index. Its core instructions are:

```
Inspect this actual indoor RGB observation. Box labels O0,
O1, ... identify concrete tracked instances, not
categories. Crops only help identity; spatial relations
must be supported by the FULL image. Never add a relation
from common sense alone. Return JSON relations with source,
target, type, confidence, endpoints_identifiable, and
visual_evidence. Allowed types: NEAR, ON, ABOVE, INSIDE,
ATTACHED_TO. NEAR requires visible local adjacency, not
merely co-visibility. ON requires support/contact; ABOVE
requires meaningful vertical alignment; INSIDE requires
containment; ATTACHED_TO requires fixed attachment. Omit
```

```
unknown edges or mark low confidence.  Do not connect an
object to an unindexed neighbor.
```

The source prompt also asks for local functional-zone annotations as evidence metadata. Those annotations do not create region nodes or spatial edges. Directionally contradictory relations observed across views are quarantined for review rather than counted twice.

C.3 Prototype descriptions and representative images

For a candidate prototype, the summarization prompt requests a concise intrinsic description based only on member evidence, separates shared structure from optional appearance, and forbids invented spatial neighbors. A separate visual audit judges representative crops with the fields `category_match`, `completeness`, `occlusion`, `clarity`, and `recognizable`. Its key instruction is:

```
Judge object-crop quality from the supplied ACTUAL images.
Complete means the main body and characteristic structural
parts are visible without substantial truncation.  Do not
mark a crop complete merely because the category is
recognizable.  Use uncertain when the evidence is ambiguous.
```

The construction pilot retained at most three distinct-instance image representatives per object prototype. This is a storage policy, not a claim that every prototype had three acceptable views. The prototype graph can use the same quality policy when selecting the image evidence represented by an image-form node.

D Additional Qualitative Cases

The supplied qualitative figure predates the form-specific Nexus. We retain three cases whose evidence types remain applicable: an image-goal bed case, an image-goal toilet case, and a text-goal toilet case. We omit its chair/office panel because it depicts an explicit region node, which the current graph does not contain. These cases illustrate possible complementary cues; they are not additional quantitative trials of the final graph.

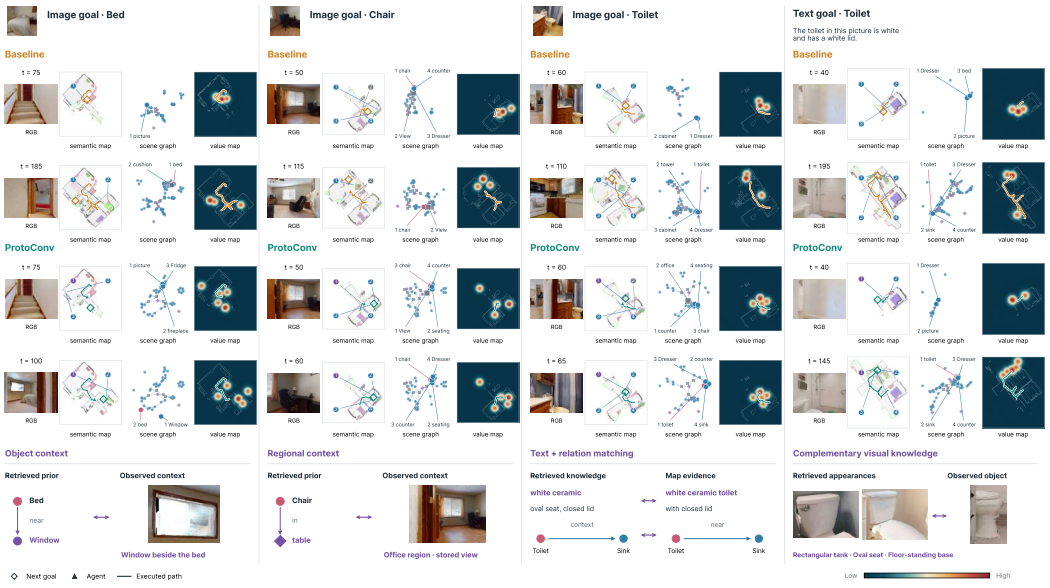

Figure 6: Image-goal bed case from the supplied qualitative figure. The observed window provides object context while the response map evolves over the trajectory. This panel is an illustrative case from the earlier visualization and is not a new performance measurement.Image-goal toilet case. Appearance and observed relations provide different cues for interpreting the local subgraph. The plot is retained as an illustrative case, not as evidence of a final graph statistic.Text-goal toilet case. Reference appearances can complement a textual goal when the online view differs in pose or visible attributes. The case is qualitative and does not establish a causal gain on its own.