2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

PersONAL: Towards a Comprehensive Benchmark for Personalized Embodied Agents

Filippo Ziliotto[1,2], Jelin Raphael Akkara[1,2], Alessandro Daniele[2], Lamberto Ballan[1], Luciano Serafini[2], Tommaso Campari[2]

***Abstract*— Recent advances in Embodied AI have enabled agents to perform increasingly complex tasks and adapt to diverse environments. However, deploying such agents in realistic human-centered scenarios, such as domestic households, remains challenging, particularly due to the difficulty of modeling individual human preferences and behaviors. In this work, we introduce PersONAL (*PERS*onalized *O*bject *N*avigation *A*nd *L*ocalization), a comprehensive benchmark designed to study personalization in Embodied AI.**

Agents must identify, retrieve, and navigate to objects associated with specific users, responding to natural-language queries such as *find Lily's backpack*. PersONAL comprises over 2,000 high-quality episodes across 30+ photorealistic homes from the HM3D dataset. Each episode includes a natural-language scene description with explicit associations between objects and their owners, requiring agents to reason over user-specific semantics.

The benchmark supports two evaluation modes: (1) active navigation in unseen environments, and (2) object grounding in previously mapped scenes. Experiments with state-of-the-art baselines reveal a substantial gap to human performance, highlighting the need for embodied agents capable of perceiving, reasoning, and memorizing over personalized information; paving the way towards real-world assistive robot. Code and dataset available at: github.io/PersONAL

***Index Terms*— Embodied AI, Personalized Agents, Benchmarks**

I. INTRODUCTION

In recent years, Embodied AI has significantly advanced, enabling agents to perform complex tasks and interact more naturally with their environments. Modern methods combine end-to-end training with zero-shot capabilities powered by large language models (LLMs), allowing agents to answer dynamically to user input [1]–[7]. Yet, their application to user-centric scenarios, where agents must interpret local, implicit information not directly encoded in pretrained models, such as object ownership, remains largely unexplored. Bridging this gap is key to deploying embodied agents in real-world environments like homes or offices. While personalized vision-language models (VLMs) [8]–[10] have been developed for user-specific visual grounding, they are typically limited to static, image-based contexts. In contrast, embodied agents must operate in complex physical settings, reasoning and acting over time.

Recently, a few works have begun to explore personalization in embodied scenarios [11], [12], but the field remains in its early stages and these works mainly focus on guiding agents using image-based queries or continuous human-robot interaction, which limits scalability and real-world applicability.

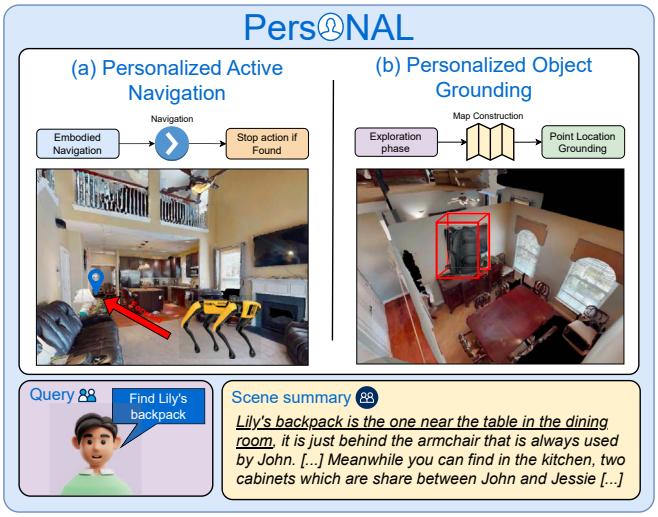

Fig. 1. We introduce **PersONAL** , a comprehensive benchmark to evaluate Embodied AI agents in the context of user centric tasks. PersONAL supports two evaluation modes: (1) active navigation in unseen environments, and (2) object grounding in previously mapped scenes.

To address this gap, we introduce PersONAL , an Embodied AI benchmark for personalized, user-centric navigation and personalized object grounding (Figure 1). Agents must interpret user-specific queries and either navigate to or retrieve the location of objects associated with particular individuals (e.g., "Navigate to Carl's backpack"). Each episode includes a textual scene description specifying object attributes and ownership (e.g., "the upper kitchen cabinet belongs to Linda"), followed by a personalized query. Unlike prior work, we also define a grounding task which acts as an embodied memory challenge, requiring agents to recall and localize targets using internal maps.

To address this gap, we introduce PersONAL , an Embodied AI benchmark for personalized, user-centric navigation and personalized object grounding (Figure 1). Agents must interpret user-specific queries and either navigate to or retrieve the location of objects associated with particular individuals (e.g., "Navigate to Carl's backpack"). Each episode includes a textual scene description specifying object attributes and ownership (e.g., "the upper kitchen cabinet belongs to Linda"), followed by a personalized query. Unlike prior work, we also define a grounding task which acts as an embodied memory challenge, requiring agents to recall and localize targets using internal maps.

[1]University of Padova, Italy. jelinraphael.akkara@phd.unipd.it, lamberto.ballan@unipd.it
[2]Fondazione Bruno Kessler (FBK), Trento, Italy. {fziliotto, daniele, serafini, tcampari}@fbk.eu

In summary, our main contributions are:

- We present PersONAL , a comprehensive Embodied AI benchmark specifically designed to evaluate embodied personalization, incorporating user-centric queries and object–ownership semantics.
- We release a dataset of 2,000 high-quality episodes sampled from over 30 realistic household environments, divided into three difficulty levels (easy, medium, and hard).
- We provide empirical analyses with state-of-the-art zero-shot baselines, showcasing limitations and possible future research directions toward human-level personalized navigation in Embodied AI.

II. RELATED WORKS

Research in embodied visual navigation has accelerated with the emergence of large-scale photorealistic simulators and datasets [13]–[16], enabling a range of navigation tasks including point-goal, object-goal, image-goal, and language-guided navigation [15], [17]–[20]. Therefore, we review prior work in embodied AI, with a particular focus on navigation tasks, and discuss related research on personalization within computer vision.

A. Embodied Navigation

Progress in embodied navigation has been driven by advanced simulation platforms and rich 3D datasets. AI2-THOR [14] and Habitat [13] provide efficient, photorealistic environments supporting large-scale training and evaluation of navigation, manipulation, and visual QA tasks. Datasets like Matterport3D, Gibson, and the larger, more diverse HM3D [15] enable agents to learn and generalize in realistic settings. These resources support reproducible research in ObjectNav, ImageNav, and grounding tasks [18], [20], [21].

Recent benchmarks have further increased realism and complexity; for example, e.g. the Open-Vocabulary ObjectNav challenge [22] uses HM3D with 15k object instances across 379 categories and free-form text goals. GOAT-Bench [23] introduces multimodal target sequences. On the other hand, the FindingDory Benchmark [24] evaluates the long-term memory capabilities of embodied agents, requiring tasks such as"navigate to the first room you visited" after exploration. However, all these works lack the focus on personalization capabilities that embodied agents require to function effectively in real-world households or office environments.

B. User-centric AI Agents

Prior work on personalized vision-language models (VLMs) [8]–[10] has focused on user-specific visual grounding in static, image-based settings. However, personalization in embodied AI—where agents must interpret and act on user queries in dynamic physical environments—remains largely unexplored. Our work addresses this gap by extending personalization to real-world scenarios.

Standard Embodied AI typically treats targets as generic instances (e.g., any chair), without distinguishing ownership [17], [20], [23]. Recent work has begun to address this

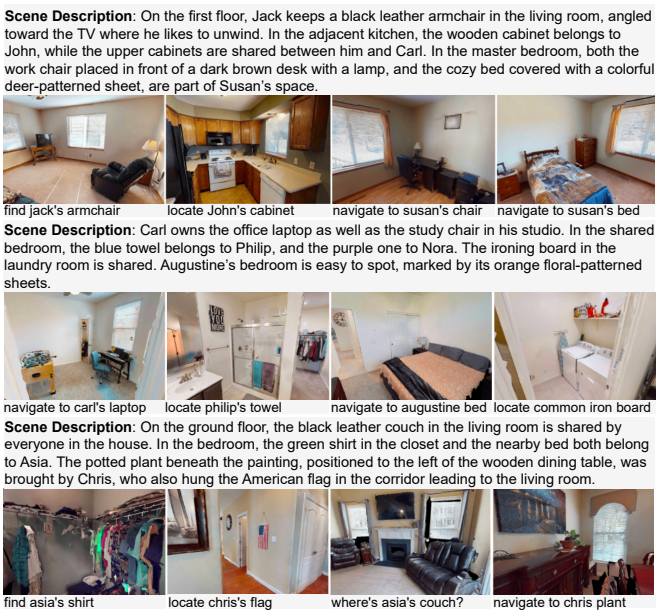

Fig. 2. **Example episodes from PersONAL .** The agent receives, as input, a personalized query along with a scene description specifying ownership details. Both navigation and grounding tasks are conditioned on this information.

by introducing Personalized Instance-based Navigation [11], where agents must locate specific user-owned objects among multiple similar items inputted as images.

Other works have incorporated personalization in object-goal navigation [12], but these benchmarks are not rigorously defined and depend on continuous user interaction through LLM-based API calls, which are impractical for real-world deployment. Furthermore, such queries do not account for scenarios involving multiple objects owned by different individuals. Instead, we advocate for approaches like [25], where agents can remember user-centric habits and learn about the environment over time without external dependencies.

User-centric embodied AI requires adapting agent behavior to individual preferences (e.g., caution), yet personalization remains largely underexplored and is still in its early stages despite recent multi-objective RL approaches [26]. In this work, we present PersONAL , a robust benchmark for Embodied AI aimed at advancing research in this area.

III. BENCHMARK

In this section, we introduce the proposed PersONAL benchmark and its associated tasks. The benchmark is structured into two main components: (1) Personalized Active Navigation (PAN), where the agent must navigate to a user-specified object without any prior knowledge of the environment; and (2) Personalized Object Grounding (POG), where the agent can pre-explore the scene to build a representation of the environment (e.g., a map) to assist in localizing the target afterwards.

In both settings, the agent receives a detailed scene description specifying object ownership (e.g., "in the kitchen, the cabinet on the left belongs to Lisa, while the one on the right belongs to James. Meanwhile, the picture above the bed belongs to ..."), as well as a user query such as "Find

Lisa's cabinet." In the POG task, the query and the scene description are provided *after* the pre-exploration phase. The agent must correctly interpret both the scene description and the query to solve the task, as shown in Figure 2. In contrast to previous work [11], our approach provides the object reference exclusively in textual form, mirroring how humans typically communicate ownership or describe personal items (e.g., "my bag is the one with blue and white stripes"). We deliberately avoid relying on image-based cues, which are uncommon and often impractical in real-world scenarios, since it requires collecting images for every single object belonging to a specific user.

A. Personalized Active Navigation (PAN)

In the Personalized Active Navigation (PAN) setting, the agent is given a budget of 500 steps to reach the target location from its initial position. The agent must rely solely on the given personalized information, without any further user interaction during the episode. This constraint reflects realistic use cases—after receiving an initial description, a robot should autonomously identify and localize the target object without continuous user input confirmation.

The task settings follow standard ObjectGoal Navigation conventions [17]. Within the Habitat framework, we set the agent's camera height to 0.88 meters for both RGB and Depth sensors, with a field of view (FOV) of 79°, imitating the embodiment of a LocoBot robot. The action space consists of MOVE_FORWARD (0.25 m), TURN_LEFT and TURN_RIGHT (30° each), LOOK_UP and LOOK_DOWN (30° each), and the STOP action. An episode is considered successful if, after selecting the STOP action, the agent lies within 1.0 meters from the target object.

B. Personalized Object Grounding (POG)

In the Query Grounding task, the agent is first allowed to pre-explore the environment with a budget of at most 2,500 steps, during which it can construct a spatial representation of the environment. We note that the robot configuration in the pre-exploration phase is identical to that used in the PAN task. We place no restrictions on the map representation: it can be a compressed 2D top-down view derived from the point cloud, a full 3D reconstruction, or even a video sequence with frame-by-frame updates. We developed a zero-shot approach to address the first 2D representation, while the latter options are left for future work. Following the exploration phase, the agent is presented with an input query and is required to localize the corresponding target object by outputting its coordinates on the given map.

IV. Dataset

Here we describe the composition and generation process of the PersONAL dataset. We curated over 2,000 episodes designed to evaluate agents' ability to interpret user queries in the context of personalized human–object ownership information. To ensure the benchmark remains challenging and broadly useful, the dataset is divided into three difficulty levels: "easy", "medium", and "hard" (see Figure 4). The core

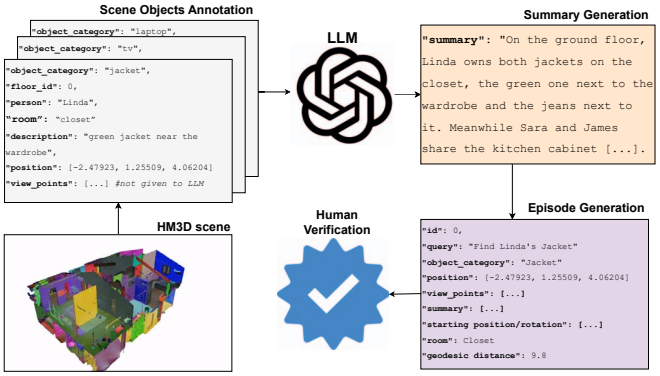

Fig. 3. **PersONAL Dataset generation**. We selected a substantial subset of episodes to manually assess the quality of the generated episodes.

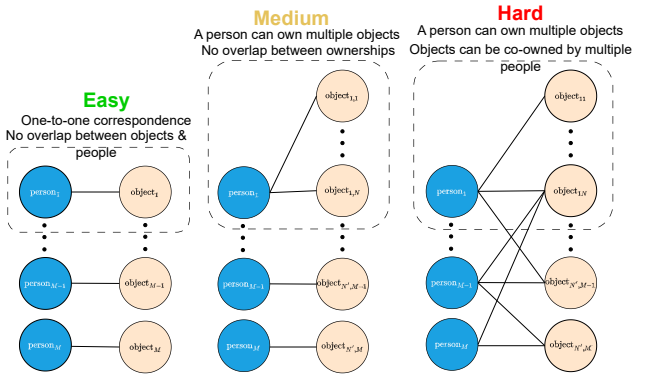

Fig. 4. **PersONAL Dataset splits**. Difficulty levels are represented as bipartite graphs between objects and owners.

idea is that personalization can, in principle, be addressed by storing user–object associations in an explicit memory or database. Or similarly, using model adaptation methods like adapters [27], [28]. However, naive memorization does not scale, and the three difficulty levels are specifically designed to encourage methods that achieve efficient personalization.

PersONAL builds on the work of GOAT-Bench dataset [23], which provides textual descriptions for objects within HM3D environments. In particular, it comprises over 140 object categories distributed across three validation splits: "seen," "seen_synonyms," and "unseen." However, during the dataset generation, we observed that a substantial fraction of these original descriptions were either non-informative or factually incorrect; specifically, we found that approximately half of the provided captions suffered from these issues.

Caption Refinement. To address this, we developed a dedicated annotation tool that enabled manual refinement of object captions. This process focused on producing more realistic, descriptive, and spatially grounded annotations (e.g., "the red backpack under the table"), in line with the overall goal of PersONAL : to introduce realistic and actionable human-centric descriptions into embodied navigation tasks. We manually refined the captions for over 3,000 individual objects, discarding non-informative or irrelevant entries to ensure high-quality and meaningful annotations throughout the dataset. As shown in Table I, our dataset exhibits greater lexical diversity and less repetition, with a more balanced

Metric	GOAT-Bench [23]	PersONAL (Ours)	Gain
Words / Description (↓)	20.65 ± 13.10	16.28 ± 6.06	-21%
Vocabulary size (↑)	1344	1564	+16%
Type–Token Ratio (↑)	0.051	0.069	+36%
Shannon entropy (bits) (↑)	6.92	7.35	+6%
Faulty Captions (↓)	40%	<10%	75%

TABLE I

DESCRIPTION QUALITY COMPARISON. GAINS ARE RELATIVE IMPROVEMENTS OF PERSONAL OVER GOAT-BENCH.

distribution of word usage compared to GOAT-Bench. In particular, Entropy increased from 6.92 to 7.35 bits (a +6% change), corresponding to a 35% increase in the effective diversity of word usage (line 3). Moreover, the average number of words per description decreased, along with the standard deviation, indicating that the descriptions are more coherent and less redundant (line 1). In contrast, GOAT-Bench often contains descriptions that are unnecessarily long or repetitive. Unlike prior works, we also labeled each object with its room and floor in an open-set manner (e.g., "the master bedroom", "the child's bed"), yielding more realistic and context-rich scene descriptions in the second phase of the dataset generation. Specific differences with respect to other benchmarks are shown in Table II.

Episode Personalization. In the second phase of dataset generation, we assigned object ownership to specific individuals for each episode, thereby establishing personalized ownership relations. Ownership assignments are modeled as a bipartite graph, with constraints applied according to the chosen split difficulty (see Figure 4). In the "easy" split each owner possesses at most one object and no overlap between object assignments, having a one-to-one correspondence. The "medium" split allows each owner to possess multiple objects, but each object can still belong to only one person. In the "hard" split, all constraints are relaxed: owners may possess multiple objects, and objects may be shared among multiple people. We note that, as shown in Table III, active navigation setting splits are further differentiated based on the agent's starting distance from the target.

In mathematical terms, let $A \in \{0,1\}^{M \times N}$ be the ownership matrix, where M is the number of people and N the number of objects. Define

$$r_i = \sum_{j=1}^{N} A_{ij}, \qquad c_j = \sum_{i=1}^{M} A_{ij},$$

where r_i is the number of objects owned by person i, and c_j is the number of people owning object j. The constraints for each difficulty split are:

$$\begin{cases} \text{Easy:} & \forall i:\ r_i = 1;\ \forall j:\ c_j = 1;\ M = N. \\ \text{Medium:} & \forall i:\ r_i \geq 1;\ \forall j:\ c_j = 1;\ \exists\, i:\ r_i > 1. \\ \text{Hard:} & \forall i:\ r_i \geq 1;\ \forall j:\ c_j \geq 1;\ \exists\ i,j : r_i > 1; c_j > 1. \end{cases}$$

Finally, to generate each episode, we input both the ownership graph and the refined object descriptions into a large language model (GPT-4.1 in our case), prompting it to produce a comprehensive scene description that explicitly encodes the ownership associations (see Figure 3). Please refer to the supplementary video for the detailed LLM input prompt. Each generated description is then paired with a corresponding query targeting a specific ownership relation. To ensure data quality, we manually reviewed a subset of 300 episodes after generation to assess the accuracy and correctness of the resulting "scene summaries," with more than 90% of the episodes correctly labeled. Summaries were labeled as correct if all owner–object attributions were coherent, object descriptions were consistent with the provided inputs, and the overall content was semantically accurate, reliable, and concise. The remaining 10%, primarily consisting of superfluous descriptive language or ambiguous owner–object attribution, were manually fixed and corrected. Moreover, in order to enhance variability across episodes, some scene descriptions include multiple instances of the same object category, which may be owned by a single individual or by multiple people, e.g. "Find one of Julia's beds". The main dataset statistics are shown in Table III.

Regarding the dataset construction for the active navigation pipeline (PAN) we generated both starting positions and valid viewpoints for each episode. However, GOAT-

Feature	PersONAL (Ours)	PInNED [11]	GOAT-Bench [23]	ZIPON [12]	InstImageNav [18]	ProcTHOR [29]	ION [30]	ZSON [31]	AI2-THOR [14]	Gibson [32]	Robo-THOR [33]	MultiON [20]
Photo-Realistic Scenes	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✗	✓
Object Caption	✓	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓
Room & Floor Annots.	✓	✗	✗	✗	✗	✓	✗	✗	✓	✓	✗	✓
Scene Description	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Object Ownership	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Navigation Task	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Grounding Task	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

TABLE II

COMPARISON OF PERSONAL WITH EXISTING EMBODIED DATASETS. FEATURES ARE SHOWN AS ROWS, DATASETS AS COLUMNS.

PersONAL Statistics	Split			Total
	Easy	Medium	Hard	
Episode Entities				
N° People	4.52	4.42	6.27	5.07
N° Objects	4.52	6.09	6.42	5.68
N° Shared Objects	0.00	0.00	4.55	1.52
N° Object Categories	51	83	119	119
Avg. Summary Length	67	117	148	110
Ownership Graph				
N° Edges	4.52	6.09	16.42	9.01
Avg. Degree per Person	1.00	1.43	2.61	1.68
Avg. Degree per Object	1.00	1.00	2.62	1.54
Density	0.23	0.24	0.42	0.30
Overlap Ratio	0.00	0.00	0.73	0.24
Navigation				
N° Episodes	600	700	720	2020
Avg. Geodesic Distance	3.12	5.23	7.56	5.30
Avg. Euclidean Distance	3.40	4.80	7.09	5.10

TABLE III

DATASET METRICS BY TASK DIFFICULTY. IT COMPRISES GRAPH STRUCTURE VARIABLES AND NAVIGATION SETTINGS ENTITIES.

Bench [23] scenes do not provide annotated viewpoints for all the considered objects. To generate them, we sample candidate positions and project a line from each to the object's boundary points (i.e. its center position). A viewpoint is retained only if the line does not intersect walls or is not heavily occluded by other objects, ensuring clear visibility of targets. We then discarded all the viewpoints distanced more than 1m to the target object center. In specific cases, we relax the 1m threshold to 1.5m, since HM3D provides only the center position of each object and not its dimensions. For larger objects, such as pictures or couches, the standard 1m constraint may exclude valid points; relaxing the threshold ensures a fair evaluation in these scenarios.

V. EXPERIMENTS

In this section, we describe the baseline approaches evaluated on the proposed PersONAL dataset. For the active navigation task, we consider established zero-shot ObjectNav methods [6], [34] that combine end-to-end reinforcement learning policies with vision–language models (VLMs) or related mechanisms to guide exploration toward promising frontiers. For the grounding task, we introduce a simple baseline that leverages frontier exploration policies in combination with a memory module constructed from VLM embedding vectors.

A. PAN Baselines

Random. As a sanity check, we first implemented a Random baseline, where the agent selects actions uniformly at random from all available frontiers and terminates after 500 steps, in order to avoid biases due to varying initial distances.

Human. To establish an upper bound on performance, we also evaluated human participants on the benchmark. Specifically, we sampled 50 episodes from each difficulty split and measured the performance of 5 human subjects (adults researchers), reporting the average as the final baseline. At each step, participants were provided with the current RGB image from the agent's FOV and the four possible agent actions. All agent's configurations follow standard ObjectNav settings.

VLFM. We tested a zero-shot approach that constructs occupancy maps from depth observations to detect exploration frontiers while using RGB observations and a pre-trained vision-language model to generate a language-grounded value map. VLFM uses this value map to prioritize frontiers most likely to contain the queried object category. For a fair comparison, we use an LLM to extract the most appropriate instance (e.g., for the query "where's Linda's clothes," the LLM identifies the related association "green jacket in the closet" as the target from the scene description). It then localizes the objects using open-set detectors [35].

OneMap. Similarly to VLFM, this zero-shot method constructs an enhanced and reusable open-vocabulary feature map for real-time object search, using a probabilistic semantic map update to reduce errors in feature extraction. This approach incorporates semantic uncertainty, enabling more informed exploration and higher path efficiency. As before, we leverage an LLM to fetch the open-set target object category.

Uni-Navid. We evaluate a VLM-based model trained for VLN and assess its ability to handle personalized navigation queries by providing a scene-level description together with the user-specific query. This preserves the VLN-style interface while enabling reasoning over global context and personalization. Unlike other baselines, we do not use an LLM to extract target categories, relying solely on the model's spatial–semantic understanding.

B. POG Baselines

In this setting, all baseline maps are constructed uniformly. The agent first explores the environment using a frontier-based policy [38], generating a 3D point cloud from depth observations, which is projected into a 2D top-down map of navigable and non-navigable areas. This spatial map is discretized into 1×1 m grid cells, each storing a semantic embedding extracted from the agent's egocentric RGB observations via a frozen vision–language model (BLIP-2 [39]). The resulting feature map has dimensions 50×50×768, providing a compact spatial memory that supports efficient retrieval for grounding. As noted earlier, we do not impose constraints on the mapping itself, since the task is solely to predict the correct real-world coordinates of the queried target object.

Random. To control for potential biases, we evaluated two random baselines: (i) an agent that selects a random point in the 2D top-down map, and (ii) an agent that always predicts the map center. The latter is included to account for cases where small maps might make the center a disproportionately strong predictor of the ground-truth object location, which we aim to avoid.

Human. The human baseline in this setting is established by providing participants with visualizations of the GLB files of the HM3D scenes. Using an automated script, we record the predicted location when the user selects the target area with a mouse click, and then compute the relevant evaluation metrics accordingly.

Query Scoring (QS). As a naive zero-shot baseline, we scored the query embeddings (encoded with a VLM encoder) directly against the feature map vectors. The same encoder was used on both sides to ensure embedding alignment. The predicted location was taken as the index of the maximum similarity in the output map and compared to the ground-truth object position. Although this approach ignores the scene description, it serves to confirm that simple query matching alone cannot solve the dataset and that no shortcut exists.

Region-gated Personalized Grounding (RPG). We developed a zero-shot method for grounding textual queries in a 2D semantic feature map, as shown in Figure 5. The input feature map is built as defined previously. We then use an LLM (GPT-4.1 in our case) to extract K descriptive phrases from each summary, corresponding to the K people present (e.g., "Clara owns a white bed"). This ensures that the final K forms are *"person x owns object y in room z"*

Method	Easy			Medium			Hard			Total		
	SR↑	SPL↑	DTG↓	SR↑	SPL↑	DTG↓	SR↑	SPL↑	DTG↓	SR↑	SPL↑	DTG↓
Human	70.21	49.20	1.26	65.22	40.05	1.40	60.13	35.31	1.32	65.19	41.52	1.33
Human *w. env. memory*†	89.00	58.12	0.84	86.00	55.00	1.10	82.00	48.20	0.98	79.67	52.44	0.97
Random	0.01	0.00	6.02	0.05	0.00	6.98	0.01	0.00	0.00	0.00	0.00	8.20
VLFM [36]	5.97	0.61	5.98	5.08	0.88	6.26	4.88	0.81	6.14	5.31	0.77	6.13
OneMap [34]	**31.17**	**24.67**	**2.86**	17.98	13.15	4.83	10.67	7.95	6.93	19.94	**15.26**	**4.27**
Uni-Navid [37]	31.03	13.65	2.93	19.51	8.23	4.70	**16.67**	6.12	6.32	**20.40**	9.33	4.65

TABLE IV

PAN RESULTS. †EVALUATIONS CONDUCTED WITH PRIOR ENVIRONMENT KNOWLEDGE, AS THE MAP WAS SHOWN BEFORE THE EPISODE START.

Each phrase is encoded into embeddings of size (K, E) using a frozen text encoder. We then compute cosine similarity scores between these embeddings and the semantic map embeddings, selecting the top-3 matching location indices for each phrase while masking out the remaining regions, yielding a masked feature map $(H_{\text{mask}}, W_{\text{mask}}, E)$.

Next, we score the embedded query vector of size $(1, E)$ against the non-masked regions of the feature map obtained in the previous step. From all candidate sub-regions, we choose the location with the highest similarity score as the final prediction, relying only on the pixel with the maximum similarity.

While the task formally requires outputting a single (x, y) coordinate, in practice a robot must physically navigate to that location. Therefore, we argue that top-k predictions (e.g., top-3) are also meaningful: the robot can attempt to reach the first candidate, and if unsuccessful, proceed to the next predicted position.

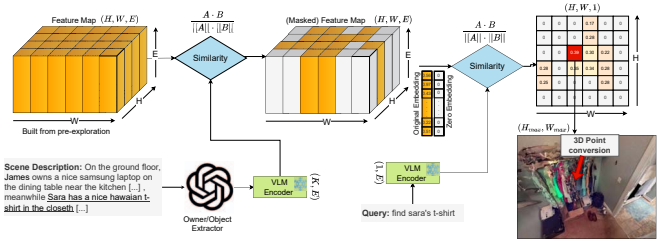

Fig. 5. **POG baseline method.** We developed a "Region-gated" method trough the use of cosine similarity for the K descriptions, selecting the top-k regions for each K, and then another query similarity to predict the 3D point in the map.

C. Metrics

In the active navigation setting, we adopt standard metrics for object-driven embodied navigation, including success rate (**SR**) and success rate weighted by path length (**SPL**). While SR simply measures whether the agent reaches the target, SPL also accounts for efficiency by penalizing unnecessarily long trajectories. In addition, we report the average distance to goal (**DTG**) at the episode's end. For the grounding task, we evaluate localization accuracy directly using SR within a 1-meter threshold, rather than IoU-based metrics, due to the lack of reliable bounding-box annotations in HM3D meshes and to better reflect realistic navigation scenarios.

D. Results

Table IV reports the results for active navigation (PAN). Human performance significantly exceeds all navigation baselines [34], [36], reaching 64% SR and up to 86% with prior environment knowledge (lines 1–2, total split). This can be attributed to humans' ability to continuously search for and selectively ground task-relevant information. Among methods, OneMap and Uni-NaVid are strongest, achieving 19.94% and 20.40% SR, respectively. OneMap is more navigation-efficient (15.26% vs. 9.33%), likely due to its LLM-based target extraction enabling open-set detection, whereas Uni-NaVid relies solely on an LLM for stop judgment. Standard VLFM yields the worst performance, likely due to poor detections from GroundingDINO [40], which lead to repeated failures.

These results highlight that the benchmark is far from saturation and provides meaningful headroom for the community. Moreover, we argue that one of the limitations lies in the open-set object detector: while the LLM can assign the correct caption, these captions are often too long or complex to be effectively recognized by the detector.

Regarding personalized query grounding (POG), results are shown in Table V. Overall, the Region-Gated method (line 5) performs best, with a SR of 25%, as it effectively suppresses irrelevant embeddings and is therefore less sensitive to noise, but still lags well behind human performance (line 1-4). This gap is clearly explained by humans' ability to interactively examine the map (e.g., rotating or zooming) before responding, which models cannot currently replicate. The main challenge lies in memory design: to support open-set queries, we store embeddings directly as "memory". This is particularly useful when the LLM extracts the object and its owner from the summary as descriptive information, which cannot be mapped to a fixed category. However this comes at the cost of accuracy in grounding the correct point. While stored graph-based memory components may help in this regard with common items, they still do not generalize to open-set personalization, leaving this as a direction for future work.

VI. REAL-WORLD TRANSFERABILITY

Although PersONAL is entirely simulation-based within the Habitat framework [13], its design targets real-world applicability. The use of photo-realistic HM3D scenes [15], along with common sensor configurations and action spaces compatible with standard robotic platforms (e.g., Spot, TurtleBot, LocoBot), facilitates direct policy transfer.

To demonstrate real-world feasibility, we evaluated PersONAL on a LoCoBot wx250s equipped with an Intel

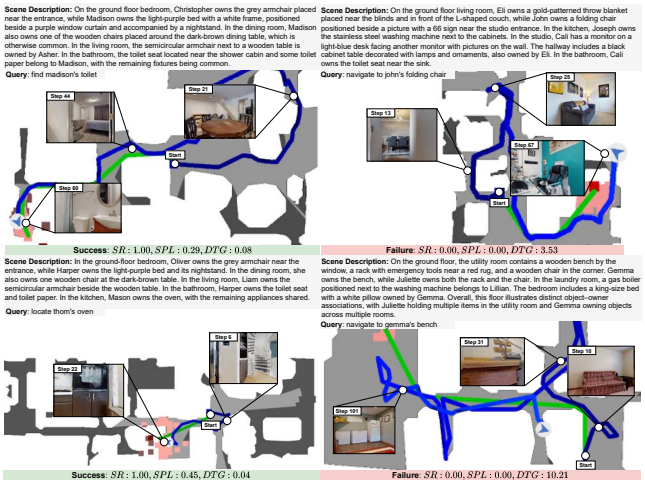

Fig. 6. **Qualitative PAN examples.** Successes (left) and failures (right) examples on the PAN active navigation benchmark episodes, using OneMap [34] (top) and VLFM [36] (bottom) baseline methods.

Map	Method	Easy SR↑	Medium SR↑	Hard SR↑	Total SR↑
GLB†	Human	91.0	89.8	88.4	89.5
2D	Random (x,y)	0.5	0.3	0.4	0.4
	Center (x,y)	1.2	0.8	1.0	1.0
	Query-Score (QS)	15.4	12.7	8.6	11.9
	Region-Gated (RPG)	**25.3**	**23.6**	**15.2**	**16.8**

TABLE V

POG RESULTS. †HUMAN SUBJECTS WERE GIVEN THE ORIGINAL GLB SCENE FILES.

RealSense D435 camera and Hector SLAM for odometry and standard ObjectNav settings. We mapped an office spanning three different office environments through a simple frontier-exploration policy until no new frontier was found (approximately 500 steps), selecting objects that were either long-tail or relatively small compared to typical office items (e.g., "art book," "black trashcan," "Amazon's bottle"). We then evaluated POG localization using the stored feature map, where our RPG algorithm achieved a 33% Success Rate (5/15), in accordance with simulated results. Each episode was manually annotated: we wrote three scene descriptions reflecting different difficulty levels, and for each description, we defined five queries.

VII. CONCLUSIONS

We introduced PersONAL a novel benchmark to tackle Personalization in embodied AI, encompassing both active navigation and query grounding scenarios, depending on whether agents possess prior knowledge of the environment. To support this benchmark, we jointly released a dataset structured into multiple difficulty levels designed to ensure long-term relevance and challenge. We evaluated several baseline approaches in a setting where agents receive detailed scene descriptions based on human-object ownership graphs, along with personalized queries referencing specific objects. Furthermore, we proposed a simple zero-shot approach that effectively addresses the grounding task in this personalized context. Our experimental results indicate that existing approaches fall significantly short of human performance. Overall, PersONAL establishes a new testbed for future research on personalized embodied AI.

Future work will target a refined PersONAL 2.0 release, including accurate ground-truth bounding boxes to support tasks such as 3D Question Answering. We also plan to integrate humans into the environments, enabling dynamic multi-target queries like "find Julia's laptop and bring it to her" to enhance practical applications.

Limitations: PersONAL serves as a strong testbed for evaluating personalized Embodied AI agents, but it also highlights limitations in current navigation pipelines. When analyzing human performance, we observe that humans achieve consistently high results once they are familiar with the environment. Even on the 'difficult' split, their performance drops only slightly compared to the easy split. This reflects the strengths of lifelong learning, where humans excel. Motivated by this, we stress the need to extend more benchmarks to support lifelong settings, enabling agents to improve navigation efficiency as they acquire experience and knowledge of the environment.

Moreover, personalization is inherently dynamic: personalized information can change over short or long timeframes (e.g., "I bought a new mug with a smiley face; I store it in my cabinet"). Current embodied datasets do not account for this variability, and a truly personalized embodied agent must be able to operate in dynamic environments.

Relying solely on LLMs is insufficient for personalization, as continuous API-based reasoning over stored information is impractical and costly. More fundamentally, future robots should unify reasoning and action rather than separate LLM-based reasoning from RL-based control, mirroring the inherently coupled nature of cognition and action in humans [41], [42].

Ethics Statement: Human involvement was limited to teleoperation (keyboard control) in simulation to establish technical baselines. No personally identifiable information, demographic attributes, or biometric data were collected. According to our institutional guidelines, this study qualifies as minimal-risk research and does not require formal Institutional Review Board (IRB) approval, as it does not involve the collection of personal data nor any intervention beyond standard computer-based interaction. Participants were informed about the nature of the experiment and provided verbal consent prior to participation. They were free to withdraw at any time.

Acknowledgements: We thank the University of Padua, Department of Mathematics "Tullio Levi-Civita", for providing the computational resources. TC and LS were supported by the PNRR project Future AI Research (FAIR - PE00000013) under the NRRP MUR program, funded by NextGenerationEU and by Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 – IPCEI-CL-0000007).

REFERENCES

[1] Y. Yang, H. Yang, J. Zhou, P. Chen, H. Zhang, Y. Du, and C. Gan, "3d-mem: 3d scene memory for embodied exploration and reasoning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 294–17 303.

[2] J. e. a. Liang, "Code as policies: Language model programs for embodied control," in *2023 IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 9493–9500.

[3] C. e. a. Song, "Llm-planner: Few-shot grounded planning for embodied agents," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 2998–3009.

[4] F. Ziliotto, T. Campari, L. Serafini, and L. Ballan, "Tango: training-free embodied ai agents for open-world tasks," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 24 603–24 613.

[5] D. Li, W. Chen, and X. Lin, "Tina: Think, interaction, and action framework for zero-shot vision language navigation," *arXiv preprint arXiv:2403.08833*, 2024.

[6] B. Wang, J. Zhang, S. Dong, I. Fang, and C. Feng, "Vlm see, robot do: Human demo video to robot action plan via vision language model," *arXiv preprint arXiv:2410.08792*, 2024.

[7] Y. Kim, D. Kim, J. Choi, J. Park, N. Oh, and D. Park, "A survey on integration of large language models with intelligent robots," *Intelligent Service Robotics*, vol. 17, no. 5, pp. 1091–1107, 2024.

[8] C. Pham, H. Phan, D. Doermann, and Y. Tian, "Plvm: A tuning-free approach for personalized large vision-language model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3632–3641.

[9] Y. Alaluf, E. Richardson, S. Tulyakov, K. Aberman, and D. Cohen-Or, "Myvlm: Personalizing vlms for user-specific queries," in *European Conference on Computer Vision*. Springer, 2024, pp. 73–91.

[10] C. Pham, H. Phan, D. Doermann, and Y. Tian, "Personalized large vision-language models," *arXiv preprint arXiv:2412.17610*, 2024.

[11] L. Barsellotti, R. Bigazzi, M. Cornia, L. Baraldi, and R. Cucchiara, "Personalized instance-based navigation toward user-specific objects in realistic environments," *Advances in Neural Information Processing Systems*, vol. 37, pp. 11 228–11 250, 2024.

[12] Y. Dai, R. Peng, S. Li, and J. Chai, "Think, act, and ask: Open-world interactive personalized robot navigation," in *2024 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2024, pp. 3296–3303.

[13] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, *et al.*, "Habitat: A platform for embodied AI research," 2019.

[14] E. Kolve, R. Mottaghi, R. Han, Y. Zhu, and A. Gupta, "Ai2-thor: An interactive 3d environment for visual ai," in *arXiv preprint arXiv:1712.05474*, 2017.

[15] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, *et al.*, "Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI," *arXiv preprint arXiv:2109.08238*, 2021.

[16] A. X. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D data in indoor environments," 2017.

[17] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "ObjectNav revisited: On evaluation of embodied agents navigating to objects," *arXiv preprint arXiv:2006.13171*, 2020.

[18] J. Krantz, S. Lee, J. Malik, D. Batra, and D. S. Chaplot, "Instance-specific image goal navigation: Training embodied agents to find object instances," *arXiv preprint arXiv:2211.15876*, 2022.

[19] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," 2018.

[20] S. Wani, S. Patel, U. Jain, A. Chang, and M. Savva, "Multion: Benchmarking semantic map memory using multi-object navigation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9700–9712, 2020.

[21] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," 2020.

[22] N. Yokoyama, R. Ramrakhya, A. Das, D. Batra, and S. Ha, "Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation," *arXiv preprint arXiv:2409.14296*, 2024.

[23] M. Khanna, R. Ramrakhya, G. Chhablani, S. Yenamandra, T. Gervet, M. Chang, Z. Kira, D. S. Chaplot, D. Batra, and R. Mottaghi, "Goat-bench: A benchmark for multi-modal lifelong navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 373–16 383.

[24] K. Yadav, Y. Ali, G. Gupta, Y. Gal, and Z. Kira, "Findingdory: A benchmark to evaluate memory in embodied agents," *arXiv preprint arXiv:2506.15635*, 2025.

[25] M. Chang, T. Gervet, M. Khanna, S. Yenamandra, D. Shah, S. Y. Min, K. Shah, C. Paxton, S. Gupta, D. Batra, *et al.*, "Goat: Go to any thing," *arXiv preprint arXiv:2311.06430*, 2023.

[26] M. Hwang, L. Weihs, C. Park, K. Lee, A. Kembhavi, and K. Ehsani, "Promptable behaviors: Personalizing multi-objective rewards from human preferences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 216–16 226.

[27] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," *Advances in neural information processing systems*, vol. 30, 2017.

[28] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[29] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, J. Salvador, K. Ehsani, W. Han, E. Kolve, A. Farhadi, A. Kembhavi, *et al.*, "Procthor: Large-scale Embodied AI using procedural generation," *arXiv preprint arXiv:2206.06994*, 2022.

[30] W. Li, X. Song, Y. Bai, S. Zhang, and S. Jiang, "Ion: Instance-level object navigation," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4343–4352.

[31] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra, "Zson: Zero-shot object-goal navigation using multimodal goal embeddings," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 340–32 352, 2022.

[32] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson Env: Real-World Perception for Embodied Agents," 2018.

[33] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford, *et al.*, "Robothor: An open simulation-to-real embodied ai platform," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3164–3174.

[34] F. L. Busch, T. Homberger, J. Ortega-Peimbert, Q. Yang, and O. Andersson, "One map to find them all: Real-time open-vocabulary mapping for zero-shot multi-object navigation," *arXiv preprint arXiv:2409.11764*, 2024.

[35] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[36] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "Vlfm: Vision-language frontier maps for zero-shot semantic navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 42–48.

[37] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, "Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks," *arXiv*, vol. 2412.06224, 2024.

[38] B. Yamauchi, "Frontier-based exploration using multiple robots," in *Proceedings of the second international conference on Autonomous agents*, 1998, pp. 47–53.

[39] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[40] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*. Springer, 2024, pp. 38–55.

[41] Z. Durante, R. Gong, B. Sarkar, N. Wake, R. Taori, P. Tang, S. Lakshmikanth, K. Schulman, A. Milstein, H. Vo, *et al.*, "An interactive agent foundation model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3652–3662.

[42] P. Fung, Y. Bachrach, A. Celikyilmaz, K. Chaudhuri, D. Chen, W. Chung, E. Dupoux, H. Gong, H. Jégou, A. Lazaric, *et al.*, "Embodied ai agents: Modeling the world," *arXiv preprint arXiv:2506.22355*, 2025.