

Perspective from a Broader Context: Can Room Style Knowledge Help Visual Floorplan Localization?

Bolei Chen, Shengsheng Yan, Yongzheng Cui, Jiayu Kang, Ping Zhong*, Jianxin Wang*

School of Computer Science and Engineering, Central South University
{boleichen, shengshengyan, yongzhengcui, jxkang, ping.zhong}@csu.edu.cn, jxwang@mail.csu.edu.cn

Abstract

Since a building’s floorplan remains consistent over time and is inherently robust to changes in visual appearance, visual Floorplan Localization (FLoc) has received increasing attention from researchers. However, as a compact and minimalist representation of the building’s layout, floorplans contain many repetitive structures (e.g., hallways and corners), thus easily result in ambiguous localization. Existing methods either pin their hopes on matching 2D structural cues in floorplans or rely on 3D geometry-constrained visual pre-trainings, ignoring the richer contextual information provided by visual images. In this paper, we suggest using broader visual scene context to empower FLoc algorithms with scene layout priors to eliminate localization uncertainty. In particular, we propose an unsupervised learning technique with clustering constraints to pre-train a room discriminator on self-collected unlabeled room images. Such a discriminator can empirically extract the hidden room type of the observed image and distinguish it from other room types. By injecting the scene context information summarized by the discriminator into an FLoc algorithm, the room style knowledge is effectively exploited to guide definite visual FLoc. We conducted sufficient comparative studies on two standard visual FLoc benchmarks. Our experiments show that our approach outperforms state-of-the-art methods and achieves significant improvements in robustness and accuracy.

Code — <https://github.com/BoLeiChen/RSKFLoc>

Introduction

Camera localization is a long-standing problem in computer vision, widely used in 3D reconstruction (Liu, Li, and Dai 2017), AR/VR applications of mobile devices, and robotic navigation (Li et al. 2024). Due to the complex room layouts and absence of satellite location signals, visual localization in indoor scenes is particularly challenging. Classical visual localization algorithms rely heavily on pre-collected databases (Balntas, Li, and Prisacariu 2018; Arandjelovic et al. 2017) or complex 3D scene reconstructions (Liu, Li, and Dai 2017; Sarlin et al. 2019; Sattler, Leibe, and Kobbelt 2016), which are expensive to build, store and maintain.

*Corresponding authors.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

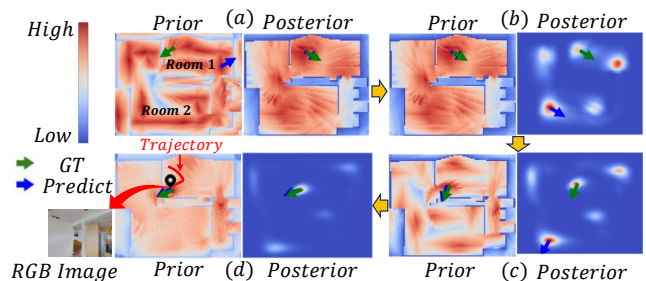


Figure 1: (a)-(d) show the ambiguous localizations in F³Loc method (Chen et al. 2024b). *Prior* denotes the probabilistic map of a single-frame visual FLoc based on the current visual image. *Posterior* denotes the probabilistic map of a long-sequence visual FLoc using Bayesian filtering. Red color indicates high likelihood.

Since floorplans are lightweight, easily accessible, consistent over time, and inherently robust to changes in visual appearance, some recent work has explored the problem of localizing camera observations by matching depth-based structural cues in the provided 2D floorplan. Such floorplans, which can be found in places such as school buildings, train stations, and apartments, encode rich and sufficient information to aid visual localization in unvisited scenes.

Existing methods (Karkus, Hsu, and Lee 2018; Chen et al. 2024b) typically match visual features with 2D geometric cues in floorplans to achieve visual Floorplan Localization (FLoc). However, as a compact representation of the building’s layout, floorplans contain many repetitive structures, thus easily result in ambiguous localization. As shown by the prior probabilistic map in Figure 1 (a), Room 1 has repetitive corner structures, leading to incorrect localization in a single room. In addition, the layouts of Room 1 and Room 2 are very similar, which can easily mislead the visual FLoc method to localize itself in the wrong room, as shown by the posterior probabilistic maps in Figure 1 (b) and (c). To address these issues, existing methods make efforts in 2D semantic/3D geometric priors exploitation (Min et al. 2022; Chen et al. 2025; Grader and Averbuch-Elor 2025) and Bayesian filtering-based sequential localization (Karkus, Hsu, and Lee 2018; Chen et al. 2024b).

On the one hand, the **State-of-The-Art** (SoTA) method

(Chen et al. 2025) injects 3D geometric priors into a visual FLoc algorithm through unsupervised learning, significantly improving FLoc accuracy. However, 3D geometric cues can not effectively resolve localization ambiguities caused by misleading scene layouts. Some other methods (Min et al. 2022; Grader and Averbuch-Elor 2025) utilize additional room category annotations or semantic labels in the floorplan (e.g., windows and doors) to assist visual FLoc. However, these semantics are not always available in floorplans or require costly manual annotations as supervised signals. On the other hand, since image sequences (Sarlin et al. 2022, 2023) can somewhat eliminate localization uncertainty, other methods (Karkus, Hsu, and Lee 2018; Chen et al. 2024b) use Bayesian filtering to optimize the posterior distribution of the current pose. However, we experimentally find that only when there is a significant change in visual appearance (e.g., when about to exit a room) can localization ambiguity be effectively alleviated, as shown by the probabilistic maps in Figure 1 (d). That is, image sequences can only mitigate FLoc ambiguity to a limited extent.

In this work, we propose a principled method to mitigate localization uncertainty caused by repetitive or similar structures by utilizing broader scene contextual information. This means that we have to figure out what we can rely on to infer indoor room relationships when only RGB images are available. We observe that different indoor rooms, such as bedrooms, bathrooms, and kitchens, typically have their own specific styles, such as decorative styles and furniture. These variations are mainly due to the different functions and requirements of each room. Therefore, visual FLoc algorithms can potentially identify room styles from current visual signals and distinguish them from other room styles.

Technically, we propose an clustering-constrained unsupervised learning technique to train a room discriminator on a self-collected unlabeled RGB image dataset. The RGB images are collected using an automated pipeline based on publicly available indoor scene datasets (Xia et al. 2018) and corresponding robotic navigation datasets (Mezghan et al. 2022), including images from different angles within the same room and images across rooms. The well-trained room discriminator can empirically extract hidden room types from observed images and explicitly distinguish them from other room types. Our visual FLoc method is built upon F³Loc (Chen et al. 2024b) and consists of a front-end observation model and a back-end histogram filter. By injecting the scene layout priors summarized by the discriminator into the observation model, the room style knowledge is effectively exploited to guide definite visual FLoc. We conducted sufficient comparative studies on two standard visual FLoc benchmarks to evaluate our method. Our method achieves SoTA visual FLoc performance and significantly outperforms the strong baselines. Overall, our main contributions are as follows:

(1) We discuss the feasibility of learning indoor room relationships from RGB images and propose a principled solution to alleviate the localization ambiguity of visual FLoc.(2) To model room style knowledge, a clustering-constrained unsupervised learning technique is proposed to train a room discriminator on an automatically collected unlabeled RGB

image dataset.(3) By injecting room style knowledge summarized by the discriminator into a visual FLoc algorithm, our method achieves SoTA FLoc performance.

Related Work

Floorplan Localization

FLoc tasks are often associated with LiDAR-based Monte Carlo Localization (MCL) (Dellaert et al. 1999; Chu, Kim, and Chen 2015; Mendez et al. 2018; Winterhalter et al. 2015), which is a classical framework for 2D localization on purely geometric maps. However, the usage of LiDAR hinders the application of such localization algorithms on common mobile devices. To alleviate this limitation, some work (Boniardi et al. 2019; Chu, Kim, and Chen 2015; Howard-Jenkins and Prisacariu 2022; Howard-Jenkins, Ruiz-Sarmiento, and Prisacariu 2021; Min et al. 2022) investigate visual FLoc based on monocular and panoramic images. Some of these methods leverage 2D scene priors (Boniardi et al. 2019) and visual features (Min et al. 2022) by matching them with scene layouts to achieve visual FLoc. Several other methods (Howard-Jenkins and Prisacariu 2022; Howard-Jenkins, Ruiz-Sarmiento, and Prisacariu 2021) localize by comparing the panoramic image features rendered at specific locations with the query image features. However, these methods either assume known camera and room heights or require panoramic images, which limits the generalization of the localization algorithms.

Recently, researchers have been working on generic monocular vision FLoc techniques (Karkus, Hsu, and Lee 2018; Chen et al. 2024b, 2025) that employ Bayesian filters (Jonschkowski and Brock 2016; Bishop, Welch et al. 2001) to solve the long-sequence FLoc problem. Despite promising progress, these methods suffer from localization uncertainty caused by repetitive structures in floorplans. To alleviate this issue, some methods (Min et al. 2022; Mendez et al. 2020; Grader and Averbuch-Elor 2025) utilize additional semantic information or room category annotations to assist visual FLoc. However, such semantic information requires complex manual annotation and is thus not always available. In this work, we propose an unsupervised learning technique to model scene context information, which is integrated into the visual FLoc algorithm to mitigate localization ambiguity.

Unsupervised Visual Pre-training

In addition to visual pre-trainings (Chen et al. 2020; Du, Gan, and Isola 2021) based solely on RGB images, more and more work has proposed cross-modal pre-training methods (Hong et al. 2023; Chen et al. 2024a; Zhu et al. 2024) to model scene semantics and geometric priors. For example, Ego²-Map (Hong et al. 2023) proposes to learn scene priors by aligning egocentric views with 2D semantic maps in a cross-modal manner. 3DLFVG (Zhang, Luo, and Lei 2024) achieves visual grounding by aligning the 2D geometric relations in RGB images with the spatial relations between 3D objects in the point cloud. Several other 2D-3D cross-modal methods (Arsomngern, Nutanong, and Suwajanakorn 2023; Chen et al. 2022, 2024a) inject 3D geometric priors into the

2D visual models by semantically or spatially aligning RGB image features with the matched point cloud sets. Although these scene priors are demonstrated to improve the accuracy of visual FLoc (Chen et al. 2025), they fail to address the ambiguous localization at the scene level. In this paper, an unsupervised visual pre-training is proposed to model room style knowledge and provide broader scene-level contextual information for visual FLoc.

Preliminaries

Problem Definition

This work aims to localize RGB images to specific imaging locations in a 2D floorplan map F , which is represented as a matrix of dimensions $H \times W$. The floorplan is a minimalist representation of a building’s layout, which retains necessary geometric occupancy information but no semantic categories. Given a time-varying image sequence $\mathcal{I} = \{\mathbf{I}_r \in \mathbb{R}^{h \times w \times 3} | r \in \{t - k, \dots, t\}\}$ containing $k + 1$ RGB images, our objective is to predict the camera’s 2D location (x, y) and orientation angle θ at which the image \mathbf{I}_t was captured. That is, given the observation $O_{\mathcal{I}, F} = (\mathcal{I}, F)$, our goal is to infer the location parameters $S_{\mathcal{I}, F} = (x, y, \theta)$. To this end, we adopt a probabilistic framework by modeling the distribution $p(S_{\mathcal{I}, F} | O_{\mathcal{I}, F})$. We discretize the camera pose space as $\mathcal{S} = \{S_i\}$ and define a probabilistic map $P \in \mathbb{R}^{\hat{H} \times \hat{W} \times O}$ where each element $P(S_i)$ represents the probability $p(S_i | O_{\mathcal{I}, F})$ for a candidate pose S_i . Here, \hat{H} and \hat{W} denote the number of discretized cells in the x and y dimensions, respectively. O represents the number of orientation bins. The predicted camera pose is then given by:

$$\hat{S}_{\mathcal{I}, F} = \arg \max_{S_i \in \mathcal{S}} p(S_i | O_{\mathcal{I}, F}) \quad (1)$$

In this work, we investigate visual FLoc in two cases: **(1)** single-frame FLoc using only the current RGB observation \mathbf{I}_t , i.e., $k = 0$, and **(2)** using sequential multi-frame RGB images \mathcal{I} to mitigate single-frame FLoc’s uncertainty and ambiguity caused by repetitive structures.

Background: F³Loc Framework

Our work builds upon F³Loc (Chen et al. 2024b), a classic framework that estimates depth rays to perform visual FLoc given a single image or an image sequence. F³Loc consists of a front-end observation model and a back-end Bayesian filter (Jonschkowski and Brock 2016). We briefly summarize several key components of their work to provide background for our framework.

Observation Model. For both single-frame and multi-frame visual FLoc, F³Loc uses a visual encoder F_θ and an attention (Vaswani 2017) based network to learn the probability distribution of planar depth over a range of depth hypotheses. Specifically, given a query image or an image sequence, the observation model estimates per-column depth values that capture the distance from the camera to the nearest wall along specific angles. These values are then linearly interpolated to produce a fixed set of equiangular depth rays $\hat{r}_d \in \mathbb{R}^l$ that represent the floorplan depth, with l denoting the number of predicted rays. The predicted depth rays are

compared with the Ground Truth (GT) rays to calculate the likelihood scores for each grid cell and orientation, resulting in a probabilistic map $P_d \in [0, 1]^{[\hat{H}, \hat{W}, O]}$. For each candidate location (x, y) on the floorplan and each discrete orientation θ , the corresponding GT rays are generated based on the floorplan’s geometry.

To unify the single-frame and multi-frame settings in one framework, a selection network implemented as a multilayer perceptron is adopted to learn a weight $0 \leq \omega \leq 1$ from the two predictions for adaptive selection:

$$\mathbf{P}_{fuse} = \omega \text{Upsample}(\mathbf{P}_{single}) + (1 - \omega) \mathbf{P}_{mv}. \quad (2)$$

\mathbf{P}_{single} and \mathbf{P}_{mv} denote the probability distributions of planar depth from a single frame and multiple frames, respectively. The upsampling operation is used to align the dimensions. The expectation of \mathbf{P}_{fuse} provides the final prediction of 2D rays. ω is manually specified as 1 and 0 implying that only single-frame and multi-frame visual FLoc are used, respectively.

Histogram Filter. F³Loc employs a histogram filter (Jonschkowski and Brock 2016) to keep track of the localization posterior over the entire floorplan. Such a filtering scheme is particularly effective in dealing with long-sequence FLoc.

Methodology

In this section, we first present how to pre-train a room discriminator on self-collected unlabeled room images using an unsupervised learning technique. Then, we describe how the scene layout priors summarized by the discriminator are injected into the observation model to guide visual FLoc.

Room Style Knowledge Pre-training

Data Collection. To recognize the room type in a visual image and distinguish it from other room types, we need to train a room discriminator using images taken in different rooms. To avoid focusing on the variation of objects between rooms instead of the style of the room when summarizing the room relationships, it is also necessary to take images from different angles within the same room. By doing so, these collected images will have completely different objects while still representing the same room. In addition, images taken in different scenes should be available to ensure the universality of the room’s stylistic representation.

Considering the scarcity of indoor scene datasets annotated with room types, we propose an unsupervised learning technique to summarize scene layout priors from unlabeled scene images. We take the Gibson indoor scene dataset (Xia et al. 2018) as an example and use the corresponding robotic navigation dataset (Mezghan et al. 2022) to collect RGB images across rooms. Specifically, for a given navigation episode E_m from the training split, we first obtain the start position p_s and the goal position p_g , as shown in Figure 2 (left). Then, we place a robot equipped with an egocentric RGB camera at each of these two positions to capture images from different angles. Finally, we assign three category attributes to each collected image I_i to record which scene it comes from ($Scene_i$), which navigation episode it comes from (E_i), and the difficulty of the navigation episode

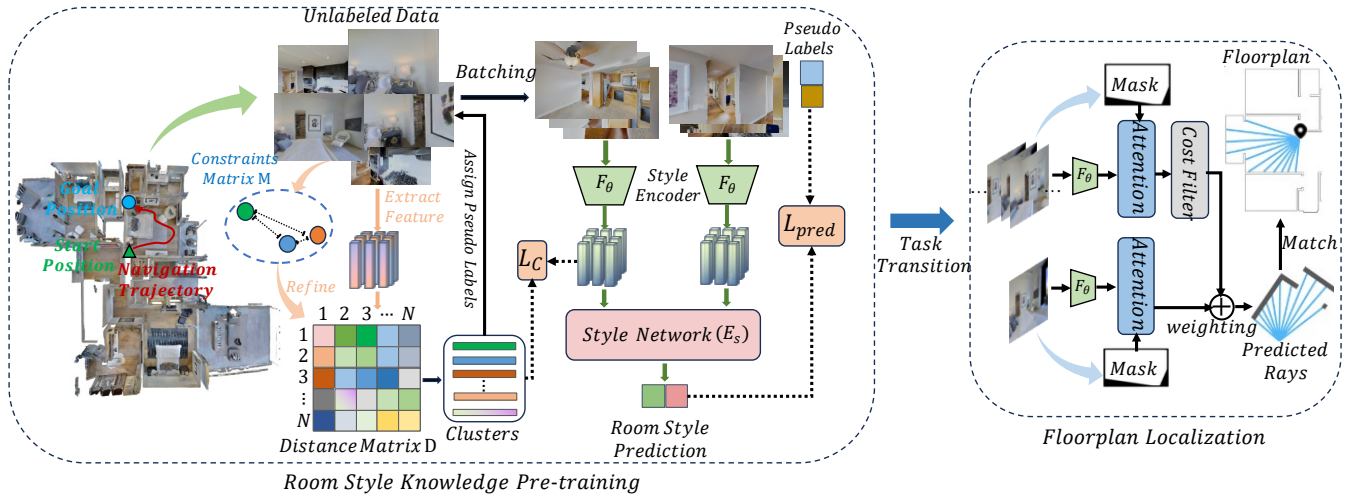


Figure 2: The left side shows the unsupervised learning technique used for room style knowledge pre-training. The unlabeled RGB images are collected automatically from indoor scene datasets based on corresponding robotic navigation episodes. We build a distance matrix D based on the cosine similarity between visual features, which is further refined by a constraint matrix M constructed based on the difficulty of the navigation episodes. Based on the refined distance matrix, the InfoMap clustering algorithm is used to cluster similar visual features and assign pseudo-labels to RGB images. The room discriminator is trained to predict room styles by optimizing a cluster-level contrastive loss L_C and a cross-entropy loss L_{pred} . The well-trained room style encoder F_θ is transferred to the right side FLoc task for fine-tuning to further fit this task.

it belongs to (E_{d_i}) . The difficulty is positively related to the length of the trajectory traversed by the navigation episode.

During the image collection process, we find that some images may contain little room style information, e.g., when the RGB camera is too close to a wall, the captured image will be completely black or white. Since using these blank images during training would provide confusing guidance, we chose to discard them. Specifically, we feed the collected images into the Segmentation Anything Model (SAM) (Kirillov et al. 2023) to obtain object segmentation masks for the whole image. If the number of object masks is less than a specified threshold, the corresponding image is considered as a blank image and discarded. Overall, our data collection pipeline is automated and can provide semantic-rich RGB images for unsupervised pre-training of room style representations.

Unsupervised Learning with Constraints. Since the data collected from the Gibson dataset lacks room annotations, we use an unsupervised training learning technique based on a clustering algorithm to train the room discriminator, which consists of a room style encoder and a style network. We observe that the navigation episodes in the Gibson dataset can be divided into three different difficulty levels according to the trajectory length: easy (1.5-3 m), medium (3-5 m) and hard (5-10 m). Intuitively, if the start and goal positions are far apart (in a hard episode), they are likely to be in different rooms. With this assumption, we use the following rules to construct a constraint matrix M of size $N \times N$ to summarize the room relationship between any two images:

(1) If $Scene_i \neq Scene_j$, then images I_i and I_j are not in the same room, set $M_{i,j} = -1$.

(2) If images I_i and I_j are taken at the same position, they are definitely in the same room, set $M_{i,j} = 1$.

(3) If $E_i = E_j$ and $Ed_i = Ed_j = easy$, then images I_i and I_j are probably in the same room, set $M_{i,j} = 0.5$.

(4) If $E_i = E_j$ and $Ed_i = Ed_j = hard$, then images I_i and I_j are probably in different rooms, set $M_{i,j} = -0.5$.

As shown in Figure 2 (left), we use the above room relation constraints to achieve room style knowledge pre-training. Specifically, a ResNet50 pre-trained on ImageNet (Deng et al. 2009) is first used as the room style encoder to extract room images as feature vectors. We then employ the cosine similarity to measure the distance between pairs of feature vectors to construct a distance matrix D . The pre-built constraint matrix M is used to refine the distance matrix D according to the following rule:

$$RefinedMatrix = D - \lambda M, \quad (3)$$

where λ is a hyperparameter. Finally, based on the refined distance matrix, we use InfoMap clustering algorithm (Rosvall and Bergstrom 2008) to cluster similar features and further assign pseudo-labels to collected RGB images. We optimize the room style encoder F_θ using a cluster-level contrastive loss, which is formulated as:

$$L_C = -\log \frac{\exp(F_\theta(I_i) \cdot \phi_+ / \tau)}{\sum_{k=1}^K \exp(F_\theta(I_i) \cdot \phi_k / \tau)}, \quad (4)$$

where K denotes the number of cluster representations and ϕ_k denotes the cluster centroid defined by the mean feature vectors of each cluster. ϕ_+ denotes a cluster center which shares the same label with image I_i . As shown in Figure 2 (left), images I_i and I_j are encoded through F_θ and then fed into a style network E_s to predict whether they are taken in

the same room. We use the cross-entropy loss to optimize the style network and style encoder, which is formulated as follows:

$$L_{\text{pred}} = - \sum_{n=1}^N [y_i \cdot \log(E_s(F_\theta(I_i), F_\theta(I_j))) + (1 - y_i) \cdot \log(1 - E_s(F_\theta(I_i), F_\theta(I_j)))] \quad (5)$$

where y_i is the assigned pseudo-labels. The overall total loss can be formulated as follows:

$$L_{\text{loss}} = L_C + \gamma L_{\text{pred}}, \quad (6)$$

where γ is a hyperparameter used to balance the two losses.

Room Style Knowledge Enhanced Visual FLoc

Through the above room style knowledge pre-training, we implicitly integrate room style knowledge into the 2D visual encoder F_θ . As shown in Figure 2 (right), the fully pre-trained room style encoder is transferred to the visual FLoc task for fine-tuning to fit the task. As described in the background section, we investigate the single-frame and multi-frame visual FLoc techniques in the F³Loc framework (Chen et al. 2024b). F³Loc localizes by finding the pose in the floorplan that has the most similar 2D rays (similar to LiDAR scans) as the prediction, as shown in Figure 2 (right). For the training of FLoc models, we optimize an L1 loss and a cosine similarity-based shape loss:

$$\mathcal{L}_{\text{FLoc}} = \|\mathbf{d}, \mathbf{d}^*\|_1 + \frac{\mathbf{d}^\top \mathbf{d}^*}{\max\{\|\mathbf{d}\|_2, \|\mathbf{d}^*\|_2, \epsilon\}}. \quad (7)$$

Where \mathbf{d} and \mathbf{d}^* are predicted and GT 2D-ray depths, respectively. ϵ is a small constant to prevent division by zero.

Experiments

Experimental Setup

Datasets. We first employ a series of Gibson (Xia et al. 2018) datasets (Gibson(g), Gibson(f), and Gibson(t)) collected by F³Loc to fully evaluate our single-frame ($Ours_s$), multi-frame ($Ours_m$), and adaptive ($Ours_f$) visual FLoc methods. We follow the data split in F³Loc, including 108 training scenes, 9 validation scenes, and 9 test scenes. The horizontal **Field Of View** (FOV) of the images in the Gibson datasets is 108°. The resolution of the floorplan extracted from the Gibson datasets is 0.1 m. Gibson(g) consists of general motions (including in-place steering motions) and includes 49,558 pieces of sequential views, each of which contains 4 image frames. Gibson(f) consists of only forward motions and includes 24,779 pieces of sequential views, each of which likewise contains 4 image frames. Therefore, Gibson(g) is intuitively more complex and harder than Gibson(f). Gibson(t) consists of 118 pieces of long-sequence views, each of which contains 280 ~ 5152 image frames.

In addition, we use the challenging Structured3D (full) (Zheng et al. 2020) dataset to perform comparative studies between our single-frame FLoc method $Ours_s$ and the SoTA methods. Structured3D (full) is a photorealistic dataset containing 3296 fully furnished indoor environments

with in total 78,453 perspective images. Notably, we use monocular images rather than panoramic images, and the horizontal FOV of each image is 80°. The resolution of the floorplan extracted from the Structured3D (full) dataset is 0.02 m. For model training and evaluation, we use the official data splits.

Baselines. We compare our method with the following FLoc baselines, none of them using semantic labels or room category annotations. **(1) PF-net** (Karkus, Hsu, and Lee 2018) proposes a particle filter specialized for visual FLoc. Its observation model aims to learn the similarity between an image and the corresponding map patch. **(2) MCL** (Dellaert et al. 1999) is the most popular framework for 2D localization on pure geometry maps. **(3) LASER** (Min et al. 2022) represents the floorplan as a set of points and gathers the features of the visible points of each pose in the floorplan. It actively compares the rendered pose features with the query image features for visual FLoc. **(4) F³Loc** (Chen et al. 2024b) is a classic visual FLoc method that proposes a probabilistic model consisting of a ray-based observation module and a histogram filtering module. F³Loc includes three variants: single-frame (F³Loc_s), multi-frame (F³Loc_m), and adaptive (F³Loc_f) visual FLoc methods. **(5) 3DP** (Chen et al. 2025) is one of the SoTA visual FLoc methods. 3DP injects 3D geometric priors into visual FLoc which significantly improve the single-frame and multi-frame FLoc accuracy without the need of any semantic labels. 3DP also includes three variants: single-frame (3DP_s), multi-frame (3DP_m), and adaptive (3DP_f) visual FLoc methods.

Metrics. Following existing work (Chen et al. 2024b, 2025), we report recall metrics computed at localization accuracies of 0.1 m, 0.5 m, and 1 m. We also report recall for predictions with an orientation error bounded to less than 30° (with a localization accuracy of 1 m). Recall is calculated as the percentage of predictions that fall within these thresholds. For comparisons on the Gibson(t) dataset, the **Root-Mean-Square Error** (RMSE) is also employed to measure the accuracy of sequential trajectory tracking when localization is successful (RMSE(S)) and in all cases (RMSE(A)).

Implementation Details. For the room style knowledge pre-training, we only collect RGB images from the training split of the Gibson dataset (Xia et al. 2018) to prevent data leakage. The height and radius of the robot used to collect data are 1.5 m and 0.1 m, respectively. The robot has a single RGB sensor with a 90° FOV. We train the style encoder and style network for 20 epochs using the Adam optimizer (Kingma 2014) with a weight decay of 5e-4 and a batch size of 64. We set the refinement hyperparameter λ as a learnable parameter. The balance parameter γ in Equation (6) is set to 1.0.

For the fine-tuning of visual FLoc, we use an Adam optimizer with a learning rate of 0.001 for all training. For the Structured3D (full) dataset, the single-frame FLoc model $Ours_s$ is trained for 100 epochs. For the Gibson(f) and Gibson(g) datasets, $Ours_s$, $Ours_m$, and $Ours_f$ are trained on the entire training split for 100, 100, and 20 epochs, respectively. During the training of $Ours_f$, the parameters of single-frame and multi-frame FLoc methods are frozen.

Method (Venue)	Gibson(f) R@				Gibson(g) R@			
	0.1 m	0.5 m	1 m	1 m 30°	0.1 m	0.5 m	1 m	1 m 30°
PF-net(CoRL 2018)	0	2.0	6.9	1.2	1.0	1.9	5.6	1.9
MCL(ICRA 1999)	1.6	4.9	12.1	8.2	2.3	6.2	9.7	7.3
LASER(CVPR 2022)	0.4	6.7	13.0	10.4	0.7	7.0	11.8	9.5
F ³ Loc _s (CVPR 2024)	4.7	28.6	36.6	35.1	4.3	26.7	33.7	32.3
F ³ Loc _m (CVPR 2024)	13.2	40.9	45.2	43.7	9.3	27.0	31.0	29.2
F ³ Loc _f (CVPR 2024)	14.3	42.1	47.4	45.6	12.2	39.4	44.5	43.2
3DP _s (MM 2025)	5.3	33.2	39.8	38.4	9.4	37.4	43.1	41.5
3DP _m (MM 2025)	15.3	42.5	47.4	45.9	11.2	36.3	41.6	39.8
3DP _f (MM 2025)	16.0	45.2	50.0	48.7	13.7	41.5	46.4	44.5
Ours _s	8.3(+3.0 ↑)	38.5(+5.3 ↑)	45.3(+5.5 ↑)	43.6(+5.2 ↑)	10.7(+1.3 ↑)	38.4(+1.0 ↑)	44.3(+1.2 ↑)	42.4(+0.9 ↑)
Ours _m	15.6(+0.3 ↑)	44.3(+1.8 ↑)	49.5(+2.1 ↑)	47.7(+1.8 ↑)	12.8(+1.6 ↑)	37.7(+1.4 ↑)	44.6(+3.0 ↑)	41.9(+2.1 ↑)
Ours _f	16.5(+0.5 ↑)	47.3(+2.1 ↑)	51.7(+1.7 ↑)	50.0(+1.3 ↑)	14.3(+0.6 ↑)	42.6(+1.1 ↑)	48.5(+2.1 ↑)	45.9(+1.4 ↑)

Table 1: Comparative studies between our single-frame (*Ours_s*), multi-frame (*Ours_m*), and adaptive (*Ours_f*) visual FLoc methods with baselines on Gibson(f) and Gibson(g) datasets.

Method (Venue)	Gibson(t)			
	R@0.2 m	R@1 m	RMSE(S)	RMSE(A)
LASER(CVPR 2022)	-	59.5	0.39	1.96
F ³ Loc _s (CVPR 2024)	35.1	89.2	0.18	0.88
3DP _s (MM 2025)	54.1	89.2	0.16	0.75
Ours _s	67.6(+13.5 ↑)	94.6(+5.4 ↑)	0.13(-0.03 ↓)	0.51(-0.24 ↓)

Table 2: Comparative studies of long-sequence visual FLoc methods on the Gibson(t) dataset.

Method (Venue)	Structured3D (full) R@			
	0.1 m	0.5 m	1 m	1 m 30°
PF-net(CoRL 2018)	0.2	1.3	3.2	0.9
MCL(ICRA 1999)	1.3	5.2	7.8	6.4
LASER(CVPR 2022)	0.7	6.4	10.4	8.7
F ³ Loc _s (CVPR 2024)	1.5	14.6	22.4	21.3
3DP _s (MM 2025)	5.6	27.4	55.5	24.0
Ours _s	6.4(+0.8 ↑)	28.6(+1.2 ↑)	56.9(+1.4 ↑)	25.2(+1.2 ↑)
Oracle	61.0	93.8	94.9	94.6

Table 3: Comparative studies of single-frame visual FLoc methods on the Structured3D (full) dataset.

Only the selection network is trained by adopting the training paradigm in (Chen et al. 2025) to prevent biased learning. All model training is performed on 4 NVIDIA 3090 GPUs. Follow the settings in (Chen et al. 2024b), Our single-frame and multi-frame FLoc methods match the predicted 40 and 160 rays to the floorplans for localization, respectively. The cost filter in Figure 2 (right) is implemented as a UNet-style (Ronneberger, Fischer, and Brox 2015) network that converts multi-channel image features into single-channel features for ray prediction. The selection network is implemented as 3 stacked linear layers with BatchNorm (Ioffe 2015) and ReLU activation.

Comparative studies with SoTA methods

We first compare our method with the baseline and SoTA methods on the Gibson (f) and Gibson (g) datasets, the

results are shown in Table 1. Existing FLoc methods are severely affected by repetitive structures in the floorplans, which can easily lead to ambiguous or even incorrect localization. Thanks to room style knowledge, our method can effectively mitigate localization uncertainty. Quantitatively, our single-frame FLoc method *Ours_s* improves the four metrics by 3.0%, 5.3%, 5.5%, and 5.2% relative to 3DP_s on Gibson(f), respectively. Even though the Gibson (g) dataset is more challenging, *Ours_s* can still improve the FLoc accuracy. Notably, *Ours_s* achieves comparable FLoc performance on both datasets, reflecting that our method can deal with drastic visual changes caused by in-place turning. Since our method can identify and distinguish room types, the performance improvement of visual FLoc extends to multi-frame and adaptive FLoc methods. As shown in Table 1, *Ours_m* and *Ours_f* significantly improve FLoc accuracy compared to 3DP. In addition, the performance gains of *Ours_m* on the more challenging Gibson(g) are greater than those on Gibson(f). Quantitatively, our multi-frame FLoc method *Ours_m* improves the four metrics by 1.6%, 1.4%, 3.0%, and 2.1% relative to 3DP_m on Gibson(g), respectively. As expected, the adaptive method *Ours_f* achieves the best performance on both datasets by integrating the strengths of single-frame and multi-frame FLoc methods.

We employ well-trained visual FLoc models to solve the long-sequence trajectory tracking problem on the Gibson(t) dataset, the results are shown in Table 2. Technically, we combine the histogram filter proposed by F³Loc with our room style Knowledge enhanced observation model. Notably, our single-frame FLoc method improves the recall metric by 13.5% compared to 3DP_s at a localization accuracy of 0.2 m. In addition, the reduction in the RMSE metrics reflects the robustness of our method in sequential trajectory tracking. In addition, we conduct comparative studies between *Ours_s* and baselines on the Structured3D (full) dataset, the results are shown in Table 3. Although Structured3D (full) dataset is challenging due to its varied scenes (across 3296 scenes), our method can still improve FLoc performance at various accuracy levels. Such performance gains imply that identifying room styles helps to clearly lo-

Ablations		Gibson(g) R@			
Data Clean	Dist Refine	0.1 m	0.5 m	1 m	1 m 30°
✓		9.4	34.7	42.1	39.5
	✓	9.8	36.8	43.0	41.2
✓	✓	10.7	38.4	44.3	42.4

Table 4: Ablation studies on data cleaning and distance refinement.

Method (Venue)	Gibson(g) R@			
	0.1 m	0.5 m	1 m	1 m 30°
SimCLR(ICML 2020)	4.7	28.2	35.3	34.6
CRL(ICCV 2021)	5.0	29.7	37.2	35.8
Ego ² -MAP(ICCV 2023)	5.7	30.6	36.9	35.2
ECL(MM 2024)	7.1	34.8	40.5	38.7
SPA(ICLR 2025)	8.3	35.7	41.4	39.5
3DP(MM 2025)	9.4	37.4	43.1	41.5
Ours	10.7	38.4	44.3	42.4

Table 5: Comparative studies of enhancing visual FLoc by using different unsupervised visual pre-trainings.

cate visual observations in floorplans. In addition, we report the FLoc performance using GT depth information (Oracle), reflecting the potential of schemes that match depth-based geometric cues. Figure 3 illustrates the changes in recall versus accuracy for long-sequence trajectory tracking using different numbers of historical frames. Our method outperforms 3DP in all localization accuracies.

Ablation Study

The ablation of room style knowledge pre-training is equivalent to rolling back our method back to F³Loc. The significant performance gains of our method relative to F³Loc highlight the effectiveness of integrating room style knowledge into visual FLoc algorithms. In addition, we perform ablation studies on data cleaning during data collection and distance refinement during pre-training, as shown in Table 4. Results on Gibson (g) indicate that both components improve FLoc performance at various localization accuracies.

As shown in Table 5, our room style knowledge pre-training is compared with existing unsupervised visual pre-trainings to demonstrate the superiority of our method. SimCLR and CRL are contrastive pre-trainings based on pure RGB images. Ego²-MAP, ECL, SPA, and 3DP are all cross-modal visual pre-trainings that use 2D/3D scene priors to enhance visual features. Technically, these comparisons are achieved by integrating their pre-trained visual encoders into the visual FLoc algorithm. The results in Table 5 emphasize the advantages of our scene context modeling compared to other visual pre-training methods.

Qualitative Evaluation

As shown in Figure 4, our method is compared qualitatively with F³Loc and 3DP in a scene with complex furniture. Compared with F³Loc, our single-frame localization is highly consistent with GT in most cases, as shown in the

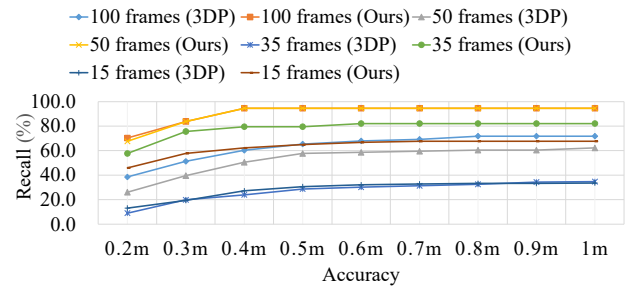


Figure 3: Comparison of the localization performance using different number of historical frames. The more frames are used within the filter, the higher the localization success.

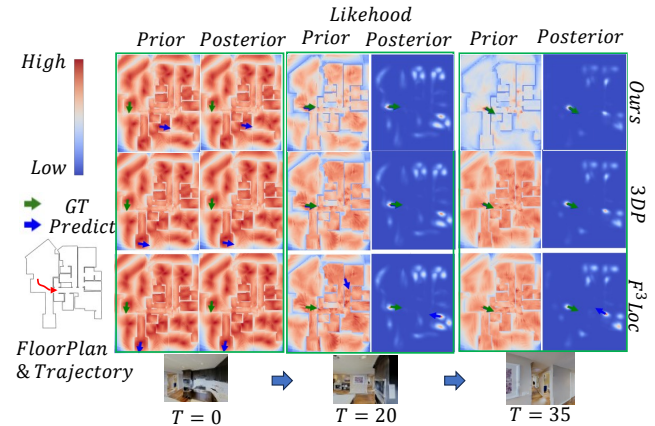


Figure 4: Qualitative comparative studies with F³Loc and 3DP in a scene with complex furniture.

prior probabilistic maps. As shown in the posterior probabilistic maps, our long-sequence trajectory tracking quickly syncs the predicted pose to the same modality with GT, whereas F³Loc shows long-term localization errors. Benefiting from our room style knowledge pre-training, the prior probability maps at $T = 20$ and $T = 35$ reflect that our single-frame FLoc method is more deterministic and can effectively alleviate localization ambiguity.

Conclusion and Future Work

In this paper, we propose using broader scene context information to mitigate ambiguous visual FLoc, rather than relying solely on 2D/3D geometric cues. In particular, we propose an unsupervised training technique based on a clustering algorithm to train a room discriminator. The room discriminator can identify and distinguish various room types, whose room style knowledge is injected into the visual FLoc algorithm to guide definite FLoc. Comparative and ablation studies on two standard visual FLoc benchmarks demonstrate the superiority of our method. Notably, this work emphasizes fully utilizing the room style knowledge for visual FLoc, which does not mean that geometric cues are not important. In future work, we will explore the advantages of integrating 3D geometric priors and scene context information by proposing a unified visual FLoc framework.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under 62272489, 62332020, and 62350004, in part by the Natural Resources Science and Technology Plan Project of Hunan Province under 2021-17, and in part by the Open Competition Project of Xiangjiang Laboratory under 23XJ01011. This work was carried out in part using computing resources at the High-Performance Computing Center of Central South University.

References

- Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2017. NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 1–1.
- Arsmongern, P.; Nutanong, S.; and Suwajanakorn, S. 2023. Learning Geometric-Aware Properties in 2D Representation Using Lightweight CAD Models, or Zero Real 3D Pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21371–21381.
- Balntas, V.; Li, S.; and Prisacariu, V. 2018. Relocnet: Continuous metric learning relocation using neural nets. In *Proceedings of the European conference on computer vision (ECCV)*, 751–767.
- Bishop, G.; Welch, G.; et al. 2001. An introduction to the kalman filter. *Proc of SIGGRAPH, Course*, 8(27599-23175): 41.
- Boniardi, F.; Valada, A.; Mohan, R.; Caselitz, T.; and Burgard, W. 2019. Robot localization in floor plans using a room layout edge extraction network. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5291–5297. IEEE.
- Chen, B.; Kang, J.; Yang, H.; Zhong, P.; and Wang, J. 2025. Perspective from a Higher Dimension: Can 3D Geometric Priors Help Visual Floorplan Localization? In *Proceedings of the 33rd ACM International Conference on Multimedia*.
- Chen, B.; Kang, J.; Zhong, P.; Liang, Y.; Sheng, Y.; and Wang, J. 2024a. Embodied Contrastive Learning with Geometric Consistency and Behavioral Awareness for Object Navigation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4776–4785.
- Chen, C.; Wang, R.; Vogel, C.; and Pollefeys, M. 2024b. F3Loc: Fusion and Filtering for Floorplan Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18029–18038.
- Chen, N.; Chu, L.; Pan, H.; Lu, Y.; and Wang, W. 2022. Self-supervised image representation learning with geometric set consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19292–19302.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.
- Chu, H.; Kim, D. K.; and Chen, T. 2015. You are here: Mimicking the human thinking process in reading floor-plans. In *Proceedings of the IEEE International Conference on Computer Vision*, 2210–2218.
- Dellaert, F.; Fox, D.; Burgard, W.; and Thrun, S. 1999. Monte carlo localization for mobile robots. In *Proceedings 1999 IEEE international conference on robotics and automation (Cat. No. 99CH36288C)*, volume 2, 1322–1328. IEEE.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- Du, Y.; Gan, C.; and Isola, P. 2021. Curious representation learning for embodied intelligence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10408–10417.
- Grader, Y.; and Averbuch-Elor, H. 2025. Supercharging Floorplan Localization with Semantic Rays. *arXiv preprint arXiv:2507.09291*.
- Hong, Y.; Zhou, Y.; Zhang, R.; Démoncourt, F.; Bui, T.; Gould, S.; and Tan, H. 2023. Learning navigational visual representations with semantic map supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3055–3067.
- Howard-Jenkins, H.; and Prisacariu, V. A. 2022. LaLaLoc++: Global floor plan comprehension for layout localisation in unvisited environments. In *European Conference on Computer Vision*, 693–709. Springer.
- Howard-Jenkins, H.; Ruiz-Sarmiento, J.-R.; and Prisacariu, V. A. 2021. Lalaloc: Latent layout localisation in dynamic, unvisited environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10107–10116.
- Ioffe, S. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.
- Jonschkowski, R.; and Brock, O. 2016. End-to-end learnable histogram filters.
- Karkus, P.; Hsu, D.; and Lee, W. S. 2018. Particle filter networks with application to visual localization. In *Conference on robot learning*, 169–178. PMLR.
- Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.
- Li, J.; Huang, W.; Wang, Z.; Liang, W.; Di, H.; and Liu, F. 2024. FloNa: Floor Plan Guided Embodied Visual Navigation. *arXiv preprint arXiv:2412.18335*.
- Liu, L.; Li, H.; and Dai, Y. 2017. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *Proceedings of the IEEE International Conference on Computer Vision*, 2372–2381.
- Mendez, O.; Hadfield, S.; Pugeault, N.; and Bowden, R. 2018. Sedar-semantic detection and ranging: Humans can localise without lidar, can robots? In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 6053–6060. IEEE.

- Mendez, O.; Hadfield, S.; Pugeault, N.; and Bowden, R. 2020. SeDAR: reading floorplans like a human—using deep learning to enable human-inspired localisation. *International Journal of Computer Vision*, 128(5): 1286–1310.
- Mezghan, L.; Sukhbaatar, S.; Lavril, T.; Maksymets, O.; Batra, D.; Bojanowski, P.; and Alahari, K. 2022. Memory-augmented reinforcement learning for image-goal navigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3316–3323. IEEE.
- Min, Z.; Khosravan, N.; Bessinger, Z.; Narayana, M.; Kang, S. B.; Dunn, E.; and Boyadzhiev, I. 2022. Laser: Latent space rendering for 2d visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11122–11131.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.
- Rosvall, M.; and Bergstrom, C. T. 2008. Maps of Random Walks on Complex Networks Reveal Community Structure. *Proceedings of the National Academy of Sciences*, 105(4): 1118–1123.
- Sarlin, P.-E.; Cadena, C.; Siegwart, R.; and Dymczyk, M. 2019. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12716–12725.
- Sarlin, P.-E.; DeTone, D.; Yang, T.-Y.; Avetisyan, A.; Straub, J.; Malisiewicz, T.; Bulo, S. R.; Newcombe, R.; Kotschieder, P.; and Balntas, V. 2023. Orienternet: Visual localization in 2d public maps with neural matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21632–21642.
- Sarlin, P.-E.; Dusmanu, M.; Schönberger, J. L.; Speciale, P.; Gruber, L.; Larsson, V.; Miksik, O.; and Pollefeys, M. 2022. Lamar: Benchmarking localization and mapping for augmented reality. In *European Conference on Computer Vision*, 686–704. Springer.
- Sattler, T.; Leibe, B.; and Kobbelt, L. 2016. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1744–1756.
- Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- Winterhalter, W.; Fleckenstein, F.; Steder, B.; Spinello, L.; and Burgard, W. 2015. Accurate indoor localization for RGB-D smartphones and tablets given 2D floor plans. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3138–3143. IEEE.
- Xia, F.; Zamir, A. R.; He, Z.; Sax, A.; Malik, J.; and Savarese, S. 2018. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9068–9079.
- Zhang, Y.; Luo, H.; and Lei, Y. 2024. Towards CLIP-driven Language-free 3D Visual Grounding via 2D-3D Relational Enhancement and Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13063–13072.
- Zheng, J.; Zhang, J.; Li, J.; Tang, R.; Gao, S.; and Zhou, Z. 2020. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 519–535. Springer.
- Zhu, H.; Yang, H.; Wang, Y.; Yang, J.; Wang, L.; and He, T. 2024. Spa: 3d spatial-awareness enables effective embodied representation. *arXiv preprint arXiv:2410.08208*.