Active Open-Vocabulary Recognition: Let Intelligent Moving Mitigate CLIP Limitations

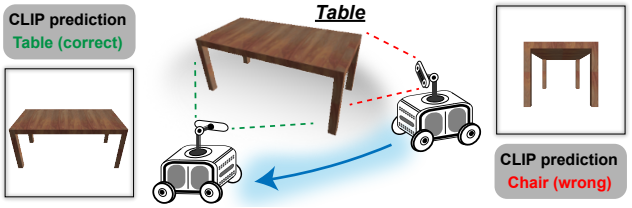
Lei Fan, Jianxiong Zhou, Xiaoying Xing and Ying Wu
Northwestern University
{leifan,jianxiongzhou2026,xiaoyingxing2026}@u.northwestern.edu,yingwu@northwestern.edu

Abstract

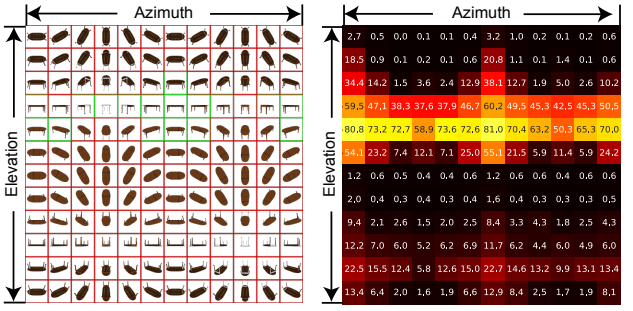
Active recognition, which allows intelligent agents to explore observations for better recognition performance, serves as a prerequisite for various embodied AI tasks, such as grasping, navigation and room arrangements. Given the evolving environment and the multitude of object classes, it is impractical to include all possible classes during the training stage. In this paper, we aim at advancing active open-vocabulary recognition, empowering embodied agents to actively perceive and classify arbitrary objects. However, directly adopting recent open-vocabulary classification models, like Contrastive Language Image Pretraining (CLIP), poses its unique challenges. Specifically, we observe that CLIP's performance is heavily affected by the viewpoint and occlusions, compromising its reliability in unconstrained embodied perception scenarios. Further, the sequential nature of observations in agent-environment interactions necessitates an effective method for integrating features that maintains discriminative strength for open-vocabulary classification. To address these issues, we introduce a novel agent for active open-vocabulary recognition. The proposed method leverages inter-frame and inter-concept similarities to navigate agent movements and to fuse features, without relying on class-specific knowledge. Compared to baseline CLIP model with 29.6% accuracy on ShapeNet dataset, the proposed agent could achieve 53.3% accuracy for open-vocabulary recognition, without any fine-tuning to the equipped CLIP model. Additional experiments conducted with the Habitat simulator further affirm the efficacy of our method.

1. Introduction

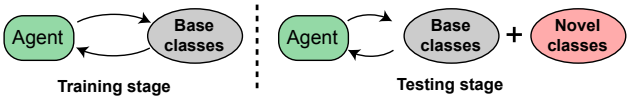
In contrast to passive visual recognition [10, 11, 20, 27, 45], where the input is obtained through human-operated or stationary devices, an active recognition agent is endowed with the ability to actively seek different observations, as illustrated in Figure 1a. This dynamic approach effectively mitigates challenges associated with undesirable viewing conditions, including ambiguous viewpoints and occlusions, by employing a sequence of intelligent movements.



(a) An illustrative episode of active recognition: An agent begins from a random viewpoint and has the freedom to execute movements to gather and aggregate information, thereby enhancing recognition performance. In this example, the integrated CLIP model fails to deliver an accurate prediction from the starting position, necessitating a change in viewpoint.



(b) The performance of CLIP on the "table" class, collected from the ShapeNet dataset. We discretize the viewing sphere surrounding each object to a 12×12 viewing grid. The heatmap reveals a significant imbalance in accuracy across various viewpoints, underscoring the importance of active observation selection in embodied agents equipped with CLIP.



(c) The task setting of active open-vocabulary recognition.

Figure 1. Illustrations of active open-vocabulary recognition task and the constraints of CLIP in embodied perception contexts.

Recent advancements have spotlighted active recognition [8, 15, 21, 44] as a critical component in a range of embodied AI tasks [24, 25, 28, 40], extending from fundamental object grasping [7] to the more intricate task of semantic goal navigation [6, 28, 31, 41, 49]. Consider the

scenario of a semantic-goal navigation agent: it must actively recognize various objects while navigating towards its target. Although numerous studies have shown promising results in active recognition, these methods predominantly operate within a closed-vocabulary framework. This limitation means that the agent is restricted to recognizing only a pre-defined set of object classes. However, the real-world environments in which embodied agents operate, such as households, encompass a vast number of object classes. This diversity renders it impractical to include all possible categories during the training phase.

Open-vocabulary recognition approaches [19, 29, 47, 48] are designed to overcome the limitations inherent in traditional models by aligning latent image-text embeddings. This allows for object classification using arbitrary text inputs during the testing phase. A notable advancement in this domain is the recent work on CLIP [9, 32, 35, 43], which elevates open-vocabulary recognition performance by collecting millions of image-text pairs and employing contrastive learning techniques. While CLIP represents a significant leap forward in single-image open-vocabulary recognition, its direct application as a visual recognizer in embodied agents presents notable challenges.

Specifically, a successful open-vocabulary recognition on agents entails three requirements: (1) Our investigations into CLIP's performance across widely-utilized platforms [5, 34, 40] for embodied AI tasks reveal a noticeable sensitivity to varying viewpoints and occlusion levels. This could undermine its reliability in embodied perception, as illustrated in Figure 1b. Consequently, an agent must possess the capability to make strategic decisions in exploring informative observations based on its current status. This requirement fits into the realm of active recognition. (2) As the agent interacts with its environment, it accumulates a sequence of observations. These observations require an effective integration mechanism to facilitate accurate class prediction, encompassing both base and novel classes. (3) The intelligent perceiving policy guiding the agent's recognition process should demonstrate robust generalization capabilities when encountering novel categories during testing.

In this paper, we introduce a novel approach for active open-vocabulary recognition, a critical yet under-explored area in embodied perception research. The task setting is depicted in Figure 1c. Our method leverages a reinforcement learning framework, enabling the agent to interact with its environment to learn the optimal recognition policy. Central to our approach is the idea of disentangling both policy and fusion method from class-specific representations. This strategy allows the model to adeptly handle novel classes encountered during testing. We postulate that recognition policies should vary based on the semantic differences between classes. For instance, the policy for recognizing a cat and a dog are likely more aligned than those for a cat and a television. Based on this intuition, our policy input is grounded in the semantic proximity between visual embeddings and the text embeddings of base categories. Moreover, we incorporate a self-attention module to allocate appropriate weights to visual features derived from a sequence of observations. These weighted features are subsequently integrated to formulate the final prediction. Our fusion method enhances performance in two significant ways: Firstly, it allows for the exclusion of inaccurate or uncertain predictions from single frames, thus preventing them from adversely affecting the integrated feature. Secondly, it could retain the discriminative efficacy towards novel classes.

The contributions of this paper can be summarized as follows: (1) We investigate the limitations of CLIP in terms of viewpoints and occlusion levels, uncovering a compelling reason to actively explore different views when employing CLIP in embodied agents. This leads to the proposition of a new task, termed active open-vocabulary recognition, aimed at enhancing embodied visual recognition. (2) To address generalization challenges, the proposed agent leverages categorical concept similarities and frame-wise similarities to guide its evidence integration and also navigation. (3) The efficacy of our agent is thoroughly evaluated across various dimensions on two widely-used platforms, namely ShapeNet [5] and Habitat [40]. Our ablation studies, focusing on different fusion strategies and policy inputs, further underscore the superiority of our method.

2. Related Work

Active recognition. Active vision, a long-standing task driven by the goal of enabling agents to intelligently acquire visual observations, has been explored through various approaches [6, 17, 18, 30, 44]. A notable branch of this field is active recognition [1–3, 8, 23] or detection [10, 16, 26], a task that empowers an agent to gather observations driven by its own motives, thereby enhancing recognition performance. Compared to the next-best-view problem [12, 42], active recognition aims for longer-term results, with its evaluation hinging on the overall movement cost.

Recent approaches in active recognition [10, 14, 21, 41] typically characterize the recognition process as a Markov Decision Process, adopting reinforcement learning to derive optimal policies. For example, in [21, 22], the authors develop an active recognition agent with an additional look-ahead module to anticipate future observations. In parallel, a more holistic approach to object understanding, termed embodied amodal recognition, is proposed in [44], offering not only categorization but also amodal bounding boxes and masks. Recently, a novel training strategy for active object detection, which combines online and offline data with a decision transformer, is introduced in [10].

However, the predominant active recognition approaches are typically limited to predefined categories. This con-

straint means that an agent is only capable of recognizing specific classes post-deployment. In [15], the authors attempt to address this by integrating continual learning into active recognition, thereby accommodating incrementally introduced classes. Despite its motivation, this approach still requires continuous training for each new class and is prone to catastrophic forgetting. In contrast, we tackle active open-vocabulary recognition in this paper, which surpasses the traditional confines of fixed categories and obviates the need for additional training after deployment.

CLIP model. The pioneering CLIP [32] and its subsequent variants [9, 35, 43] have demonstrated notable efficacy in zero-shot image recognition across various benchmarks [33, 39]. These models operate on the principle of aligning visual and textual inputs within a shared embedding space, achieved through training with extensive image-text pair datasets. In the testing phase, the model calculates cosine similarity between encoded visual features and a set of text embeddings, leading to predictions based on the highest similarity. Recent works [24, 28, 37] have integrated CLIP models, or their derivative features, into embodied AI applications, yielding substantial improvements compared to backbones pretrained on ImageNet.

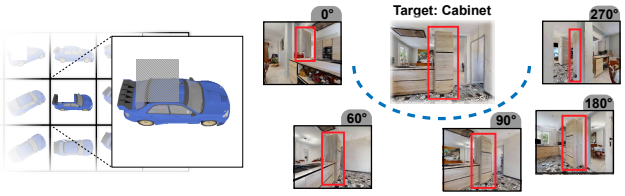
While there has been considerable exploration into both the capabilities and limitations of CLIP models from various perspectives [32, 38], studies on their performance under varying viewpoints or levels of occlusion remains sparse. Given that sub-optimal viewing conditions are common in embodied recognition scenarios, our study examines CLIP model performance under these conditions on two popular platforms [5, 40]. We find that the impact of adverse viewing conditions is significant. Consequently, for effective deployment of CLIP models as open-vocabulary recognizers on embodied agents, the ability for acquiring diverse observations, *i.e.*, active recognition, is essential.

Zero-shot object navigation. It is a task that aims to guide a robot to find a target belonging to an unseen class [28, 49]. This task diverges from active recognition; it emphasizes guiding the robot to approach the target by leveraging prior semantic relations, such as the higher likelihood of finding a mug in a kitchen rather than in a bathroom.

To conclude this section, we highlight our dual motivations in active open-vocabulary recognition, which are intrinsically synergistic. Our primary aim is to empower active recognition agents with the proficiency to effectively manage unseen classes, leveraging the capabilities of CLIP models. Concurrently, we strive to address the limitations of CLIP, such as handling viewpoint variations and occlusions, by integrating active recognition strategies.

3. When is CLIP ineffective?

Consider an embodied agent working in a household environment: the object of interest could be cluttered, creating heavy occlusions; located far from reach, resulting in an obscure sight; or relatively positioned in an ambiguous viewpoint. These undesired viewing conditions are prevalent in embodied perception scenarios because we cannot control the environment to observe or the capturing setup of the embodied agent. Similar challenges exist in related fields, such as first-person vision [13] and autonomous driving [46].



(a) Occlusions by creating random masks for ShapeNet.

(b) The collecting process of different viewpoints in Habitat.

Figure 2. The process of curating test datasets to evaluate CLIP performance with occlusions and varying viewpoints.

Consequently, it is imperative for the recognition systems used by embodied agents to adeptly handle such challenges. Before deploying CLIP in these scenarios, it is crucial to thoroughly evaluate its performance under conditions of embodied perception. Our study focuses on two key aspects of adverse viewing conditions: varied viewpoints and the presence of occlusions.

To conduct this evaluation, we select two CLIP models with Vision Transformers (ViT-B/32, ViT-L/14), one model based on ResNet-50 architecture (RN50x64) [32], and the recently developed MetaCLIP [43] for detailed examination. These models will be assessed using the zero-shot recognition approach on specifically curated datasets. Due to the constraint in page length, the main paper will primarily present the results pertaining to the ViT-B/32 model, while the analyses of the other model architectures will be detailed in our supplementary materials.

3.1. Datasets for investigation

We collect testing datasets from two widely-adopted platforms [5, 40] for testing varying viewpoints and occlusions.

ShapeNet dataset. The ShapeNetCore dataset [5] comprises approximately 41500 Computer-Aided Design (CAD) models spanning 55 common categories in their training split. This dataset is utilized for studying active object recognition [15, 21, 22], as it allows an agent to manipulate 3D objects through movements, thereby acquiring novel observations. To analyze the performance of CLIP across various viewpoints, we discretize the viewing sphere surrounding each object into 30-degree segments. This approach results in a viewing grid with $M = 12$ azimuths and $N = 12$ elevations for each object. An example of viewing grid is shown on the left side of Figure 1b. Another rationale for selecting the ShapeNet dataset is its geometric alignment within classes. All 3D models in a given class are aligned, ensur-

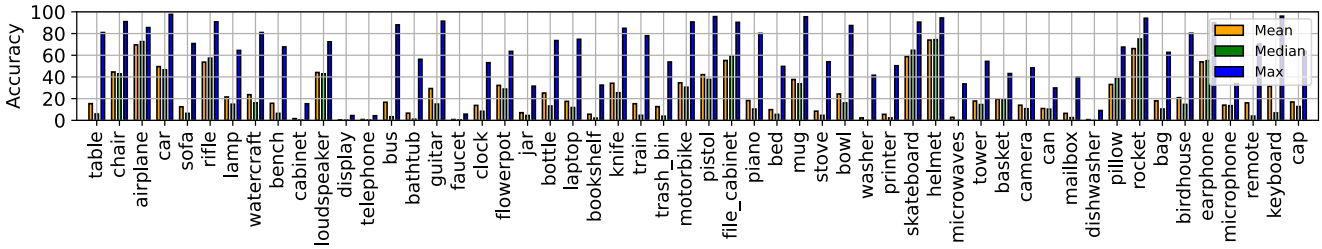


Figure 3. Performance of CLIP across all viewpoints within each category, reporting the mean, median, and maximum accuracy. The discrepancy between mean and maximum accuracy highlights CLIP's potential sensitivity to different viewpoints.

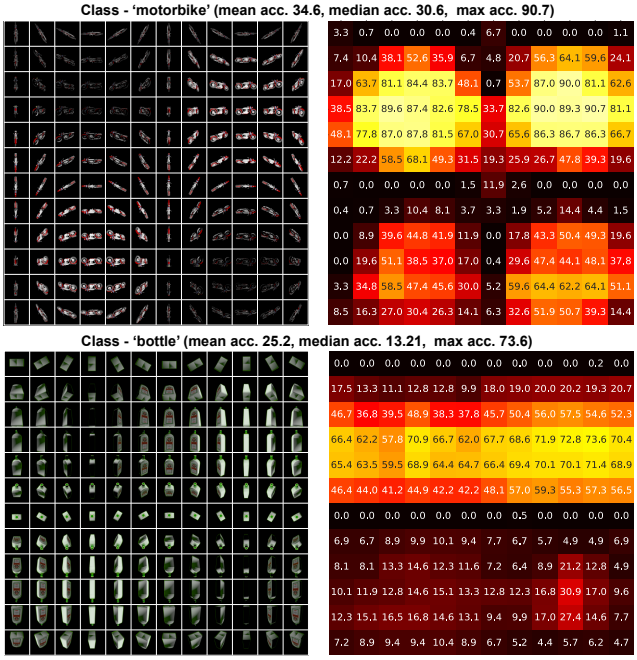


Figure 4. Recognition accuracy of different viewpoints on 'motorbike' and 'bottle' classes. The left side shows a testing sample for each class, while the right heatmap displays the average accuracy across all testing samples for each viewpoint.

ing that identical grid coordinates correspond to comparable viewpoints across different samples.

To examine the impact of occlusions, we introduce random sight-blocking patches at each viewpoint, applying them with a predetermined probability. Each patch, defined as a $\frac{1}{3}$-length square of the original view, is randomly positioned within the current visual field. An example of a contaminated viewing grid is displayed in Figure 2a. It is important to note that the level of occlusion intensifies as the probability of view-blocking increases.

Habitat dataset. To examine the efficacy of CLIP in complex indoor environments, we construct a testing dataset using the existing simulator [40] with 145 semantic annotated scenes. From 40 labeled categories, we select 25, excluding those that are ambiguous or related to building components such as doors, walls, and ceilings. This approach results in a dataset comprising 4659 objects. Detailed class distribution information is available in our supplementary materials. For each chosen object, the agent is positioned at a random location, facing the target within a maximum distance of 3.0 meters. Subsequently, the agent rotates in 30-degree increments around the target on the horizontal plane. It is important to note that images where the target is not visible are excluded from the test. The image capture process is generally illustrated in Figure 2b.

3.2. Sensitivity of CLIP to viewpoints

We first examine the sensitivity of CLIP to viewpoints, utilizing the ShapeNet dataset. We use all 12×12 views of each sample to test the CLIP's recognition with text inputs describing all categories, akin to zero-shot recognition. Given that samples within the same category are aligned in 3D, we compute the average accuracy for each specific viewpoint across all samples of the same class. This process results in a detailed accuracy map for all viewpoints. Examples are given in Figure 4. Subsequently, we calculate the mean, median, and maximum accuracy across all viewpoints, with the results presented in Figure 3.

Ideally, a proper visual recognition system should exhibit robustness to changes in viewpoint, maintaining consistent performance across different angles. However, our analysis of CLIP's performance reveals significant variations in recognition accuracy depending on the viewpoint of the same object. Specifically, the average difference between the mean and maximum accuracy across all viewpoints and classes is a substantial 40.1%, a noteworthy level considering the overall average accuracy stood at 23.8%. Consequently, this suggests that an embodied agent using CLIP and adopting a random viewing position could experience a significant decline in performance compared to an optimal viewpoint.

Further exploration is conducted using the collected Habitat dataset, which features multi-view object images from an indoor simulator. Unlike the ShapeNet dataset, images in the Habitat dataset are captured at the horizontal plane and include natural occlusions typical of indoor settings. Since objects within the same class in the Habitat dataset are not aligned, we assessed the accuracy gap on a per-sample basis. For each object, we record the highest performance across

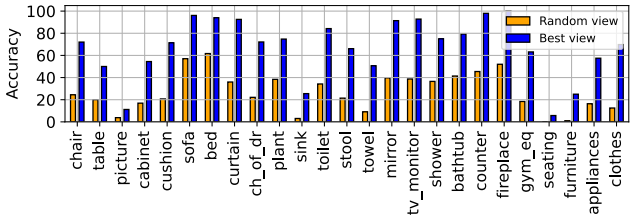


Figure 5. The accuracy comparison on the Habitat dataset between randomly taking viewpoints (average of 5 runs) and the best performing viewpoint.

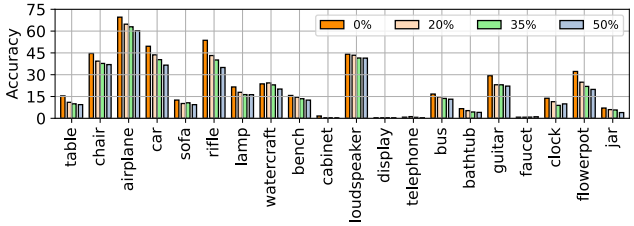


Figure 6. The mean accuracy decline observed when various occlusion levels (20%, 35%, 50%) are applied to the ShapeNet dataset.

all available viewpoints and compare it with the accuracy of a randomly selected view. The findings, depicted in Figure 5, show an average discrepancy of 40.2% between the performances of the random and optimal views, with an average random view accuracy of 26.8% across classes.

3.3. Impact of occlusions on CLIP

We conducted a preliminary experiment to assess the impact of occlusions on the performance of CLIP. Three distinct levels of random occlusions are applied to the ShapeNet dataset, as outlined in Section 3.1. Figure 6 illustrates the decrease in mean accuracy across the most common 20 classes. Notably, the average accuracy drop across all 55 classes at three different occlusion levels are 3.1%, 4.0%, and 5.0%, respectively. Although the occlusions introduced in this experiment are relatively simple compared to the more challenging occlusions encountered in embodied recognition scenarios, they nonetheless underscore the necessity of intelligent perceiving for enhancing CLIP's robustness to occlusions.

4. Active Open-Vocabulary Recognition

In this section, we introduce our proposed agent, designed to handle active open-vocabulary recognition. We outline three essential requirements for an agent to effectively manage this challenging task. Firstly, the agent must intelligently perceive novel and informative views to enhance recognition performance as discussed in Section 3. Secondly, it is important to effectively integrate accumulated evidence from observations, including for novel categories not encountered during the training phase. Thirdly, the recognition policy should also demonstrate a robust generalization capability towards novel categories. To meet these requirements, our method models the recognition process within the framework of reinforcement learning. Here, the policy is rewarded for successful recognition. Moreover, rather than recurrently aggregating single-image features for recognition [15, 21], our approach involves assigning weights to CLIP features using a self-attention module. This strategy helps to preserve the inherent capability of recognizing novel categories without adverse interference. For the policy, we incorporate a similarity measure that compares the current image with text embeddings from base classes, thereby highlighting their semantic differences. An illustrative overview of our proposed agent can be found in Figure 7.

4.1. Problem setup and notation

We outline our problem setup by initially introducing a basic active recognition agent, subsequently extending this concept to the open-vocabulary recognition scenario.

The active recognition agent is provided with a target object, denoted as x, associated with an unknown label y. The agent is permitted a total of T steps to make the final prediction $\hat{y}$ of the target. At each timestep $t = 1, \ldots, T-1$, the agent may execute an additional action $a^t \in \mathcal{A}$, signifying a movement to alter its viewing point for a new observation v_t. Here, $\mathcal{A}$ represents the predefined action space, *e.g.*, elevating the camera by 30 degrees. Through these movements, the agent aims to enhance its recognition performance by efficiently exploring its environment, aggregating observations, and classifying based on the integrated information.

Following the introduction of the active recognition task, we discuss the open-vocabulary recognition setup. During the training phase, the object presented to the agent is sampled from a set of base classes, $\mathcal{C}_B$. Conversely, in the testing phase, the target object is sourced from a broader open vocabulary, $\mathcal{C}_O$, where $\mathcal{C}_B \subset \mathcal{C}_O$. The novel classes, not included in the base classes, are collectively referred to as $\mathcal{C}_N$. Additionally, for each class within $\mathcal{C}_O$, a corresponding text embedding is available during testing, as current vision-language models, such as CLIP, necessitate a similarity measure to predict the class.

4.2. Attention-based feature integration

The image CLIP model, as described in [32], lacks the capability to aggregate temporal information while producing per-step features. For a recognition episode of the target x, we obtain a set of observations denoted as $\mathcal{V}_x = \{v_1, \ldots, v_T\}$. We omit the subscript x in subsequent formulations for clarity. The corresponding image embeddings are denoted as a sequence $\mathcal{F} = (f_1, \ldots, f_T)$, obtained from a fixed CLIP image encoder.

Each feature f could be directly utilized to predict the class of the target. However, to counter undesired viewing conditions and enhance robustness, it is critical to integrate

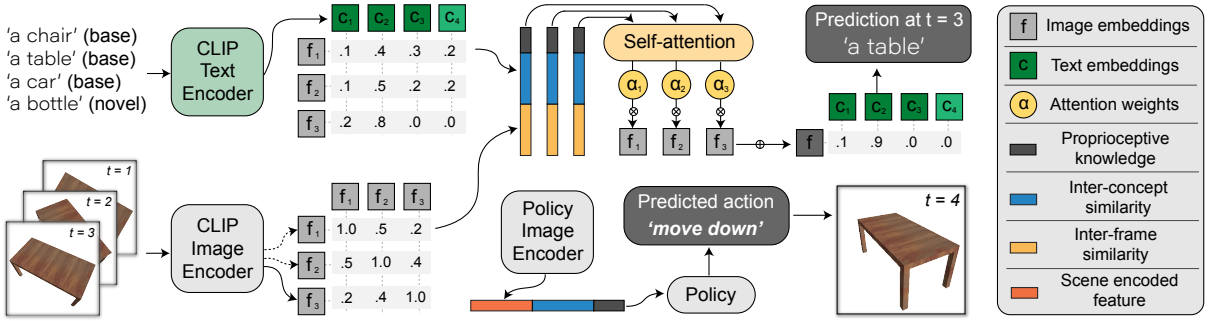


Figure 7. An illustration of the proposed architecture for the active open-vocabulary recognition agent. Only key data flows are depicted to avoid confusion. The meanings of different features and weights are detailed in the grey box on the right.

evidence across frames. We employ an attention mechanism to derive the global feature, which can be formulated as:

$$f = \mathcal{F}\alpha, \tag{1}$$

where α represents the attention weight vector of dimension T, satisfying ℓ_1-norm $= 1$. The weight α assesses the importance of each frame, filtering irrelevant information and retaining critical details in the global feature.

To obtain α, we utilize a self-attention module and a linear layer. These map input features $(q_1, \ldots, q_T)$ to a T-dimensional embedding, followed by the softmax function to finalize α. The design of the input feature q_t is crucial, as it must avoid introducing class-specific knowledge into the self-attention module. Therefore, during training, we calculate the top-k cosine similarity between the image feature f_t and text embeddings $\{c_i, i \in \mathcal{C}_B\}$, and during testing, between f_t and $\{c_i, i \in \mathcal{C}_O\}$. The process establishes the inter-concept similarity s_t^{concept}. This similarity, devoid of specific semantics, maintains a measure of recognition confidence for the current frame. Specifically, a skewed distribution of inter-concept similarity usually indicates the feature f_t contains low ambiguities. It is also important to mention that the inter-concept similarity is a preliminary approach for estimating uncertainties in CLIP features, which presumes that CLIP rarely make overconfident mis-classifications.

In addition to s_t^{concept}, we incorporate the inter-frame similarity s_t^{frame} and proprioceptive knowledge p_t into the self-attention module. The inter-frame similarity s_t^{frame} quantifies the semantic distances between the current feature f_t with features from previous frames. Meanwhile, p_t details the relative location change compared to the previous frame and the last action a_{t-1} taken. Consequently, our input feature for the self-attention module is $q_t = [s_t^{\text{concept}}, s_t^{\text{frame}}, p_t]$. During training with base classes $\mathcal{C}_B$, we apply the following loss to the global feature f:

$$\mathcal{L}_{\text{attention}} = F_{\text{cross-entropy}}(\hat{y}, y), \tag{2}$$

where $\hat{y}$ is the prediction obtained by applying the softmax function on $\{fc_i^{\intercal}, i \in \mathcal{C}_B\}$.

4.3. Active open-vocabulary recognition agent

We introduce the approach enabling the agent to intelligently navigate in order to acquire informative observations, namely, the policy module. The policy is formulated as a Partially Observable Markov Decision Process (POMDP), represented by the pdf $\pi(a_t|h_{t-1})$. Here, h_{t-1} represents the agent's prior state, recurrently derived from observations up to time step $t-1$. Specifically, the policy module integrates a single-layer Gated Recurrent Unit (GRU) to accumulate temporal information, and two linear layers function as actor and critic based on the GRU output. For the policy, we employ a separate image encoder, *i.e.*, a simple 3-layer network, to generate the scene encoded feature f_t^{policy}. This choice, instead of using CLIP features, is to preclude the infusion of strong semantic knowledge into the policy, which could potentially impede its generalization. Analogous to our feature integration module, the policy input is incorporated with inter-concept similarity $s_t^{\text{concept'}}$ and proprioceptive knowledge p_t. However, the $s_t^{\text{concept'}}$ for the policy solely focuses on the base classes $\mathcal{C}_B$, excluding the top-k operation, thus yielding a $|\mathcal{C}_B|$-dimensional embedding that reflects the current observation's relation to base classes.

The training reward for the policy is defined as the classification score of the ground-truth class y, which varies between 0 to 1. This reward changes based on the precision of predictions for the correct classes, which is derived from the global feature. We adopt the Proximal Policy Optimization (PPO) algorithm [36] to train the recognition policy from the interactions between the agent and its environment.

5. Experiment

The objectives of our experimental analysis are threefold: (1) Assessing the effectiveness of our agent in managing the open-vocabulary recognition task. (2) Evaluating the advantages of the intelligent policy in comparison to the static, single-frame CLIP model. (3) Conducting ablation studies to analyze the impact of two key components, *i.e.*, the feature integration method and the policy input.

Table 1. Recognition success rates on the ShapeNet dataset with varied class splits. The success rate is measured based on the final predictions. The difference between our method and other heuristic policies, *i.e.*, `Random`, `Largest-step`, highlights the improvements achieved through intelligent movements. `Last-prediction` represents the last-step prediction of our agent without evidence fusion.

Model	Base/novel/open classes split													
	10/45/55						20/35/55						55/0/55	
	Base classes		Novel classes		Open classes		Base classes		Novel classes		Open classes		Base classes	
	top-1	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*
CLIP (ViT-B/32) [32]	33.1	52.2	21.6	34.0	29.6	46.7	30.1	47.4	24.8	39.3	29.6	46.7	29.6	46.7
Ours + `Random`	39.3	63.1	26.5	42.5	35.4	56.9	35.4	55.8	29.0	48.9	35.4	56.9	35.4	56.9
Ours + `Largest-step`	41.0	65.6	26.8	42.5	36.7	58.7	37.8	58.2	30.1	48.2	36.7	58.7	36.7	58.7
Ours + `Last-prediction`	54.1	72.9	32.0	48.4	47.4	65.5	55.2	73.1	43.4	62.9	53.6	71.7	57.0	74.8
Ours	**60.6**	**81.3**	**36.6**	**55.1**	**53.3**	**73.4**	**57.9**	**76.8**	**47.8**	**69.0**	**56.6**	**75.7**	**59.2**	**78.8**

5.1. Datasets and experimental setup

ShapeNet. We utilized the ShapeNet [5] to train the agent, initializing the viewing grid identically to the introduced investigation dataset in Section 3.1. Starting from a random viewpoint, the agent is permitted to take a total of $T = 6$ steps. At each step, the agent can navigate within a 5×5 grid relative to its current position.

For the open-vocabulary setting, the most common 10 or 20 object categories are chosen as the base classes. The training set, comprising samples from these base classes, is available to the agent during its training stage. For evaluation, all testing samples from 55 classes are used.

Habitat. The training of agent is conducted using 145 semantically-annotated indoor scenes from HM3D [40], while testing is carried out on 90 non-overlapping scenes from MP3D [4], both datasets being part of the Habitat platform. In each training or testing episode, the agent is randomly placed in a scene with the target identified by a bounding box. The agent takes movement to view the target and predict its class label. The defined action space includes {`move_forward`, `turn_left`, `turn_right`, `look_up`, `look_down`}. Actions allow the agent to move 0.25m, turn by 10 degrees, or tilt by 10 degrees. A total of $T = 10$ steps is permitted for experiments on the Habitat platform.

We categorized the 25 classes into 10 base classes and 15 novel classes, based on their frequency in the HM3D dataset.

5.2. Result on ShapeNet

We present the comparison of our proposed agent against other baselines in Table 1. In addition to CLIP, which serves as a passive recognition baseline, we utilize two heuristic policies, namely, `Random` and `Largest-step`, to highlight the benefits of integrating intelligent movements into the embodied recognition process. The `Random` policy randomly selects an action at each step, while `Largest-step` opts for a movement most distant from the current viewpoint. `Last-prediction`, based on our agent's last observation, indicates the success of our proposed policy in locating an informative view at its final step. Except for policy alterations, all compared baselines utilize the same architecture.

We report results for two different splits between base and novel classes for open-vocabulary recognition. Specifically, either 10 or 20 base classes are introduced during training. Test samples from all 55 classes are used for evaluation. Furthermore, we present results when samples from all classes are available during training, serving as the upper bound.

Our agent achieves a 53.3% success rate when trained with only 10 classes, marking a 23.7% absolute improvement over CLIP. Additionally, the success rate rises by 15.0% when evaluating on the 45 novel classes, underscoring our agent's effectiveness in handling unseen categories. Our performance also surpasses other heuristic policies, demonstrating that random sequences of observations do not necessarily enhance recognition performance; instead, an intelligent policy is essential for effective embodied recognition.

Another finding relates to the variation in open-class results based on the number of base classes. Our method effectively learns an appropriate recognition policy even with limited base classes. For instance, when tested on open classes, our agent trained with only 10 base classes experiences just a 3.3% drop in success rate compared to training with 20 classes. This resilience likely originates from our policy and evidence integration design, which avoids class-specific representations and instead relies on similarity measures.

The step-by-step performance with 10 base classes is illustrated in Figure 8. In addition to the aforementioned baselines, we introduce another variant of our agent that predicts based solely on the current CLIP feature. Its comparison with predictions using fused global features highlights the advantages of our attention-based integration method.

5.3. Result on Habitat

Our evaluation on the Habitat simulator [40] involves allowing the agent to navigate freely within indoor environments to predict the class of a specified target. An testing episode is depicted in Figure 9. In this study, we assess the performance of our recognition agent in comparison with CLIP and other heuristic-based policies. Additionally, we introduce a new policy, `Fixation`, which is designed to maintain the target at the center of view during movement. The results of these comparisons, presented in Table 2, further validates the effectiveness of our proposed method in complex scenarios.

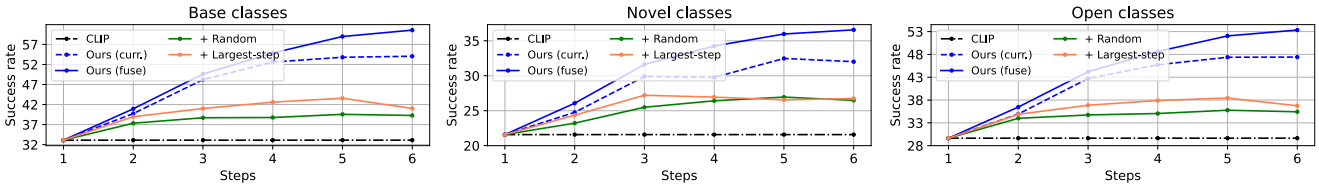


Figure 8. Performance comparison of different agents over steps on the ShapeNet dataset with the base/novel/open split of 10/45/55.

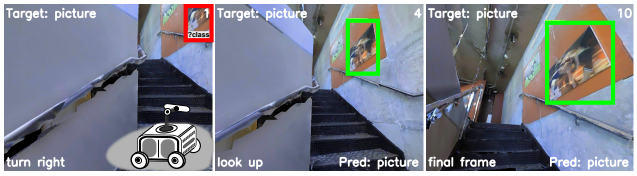


Figure 9. An episode of active recognition in the Habitat simulator. A target is queried at step $t = 1$. The agent is permitted to make 10 observations in each episode. Three steps (at $t = 1, 4, 10$) are shown, along with their subsequent actions and class predictions.

Table 2. Comparison of performance on the Habitat dataset using a class split of 10 base classes and 15 novel classes.

Method	Base classes		Novel classes		Open classes	
	top-1	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*
CLIP (ViT-B/32)	22.2	43.1	32.3	55.0	24.9	46.2
Ours + `Random`	22.3	43.3	32.5	55.2	25.1	46.4
Ours + `Fixation`	23.4	45.1	33.7	56.3	26.2	47.7
Ours	**25.8**	**48.8**	**35.4**	**58.1**	**28.0**	**49.8**

5.4. Ablation studies

In this study, we investigate the effects of various factors, such as different evidence integration strategies and the choice of input features for the policy. The experiments are conducted using the ShapeNet dataset with 10 base classes.

We assess four alternative evidence integration strategies in comparison with our proposed attention-based method: *Average-feature*, *Average-prediction*, *Max-prediction*, and *Vote*. The *Average-feature* approach computes the mean of accumulated CLIP features across frames prior to making a prediction. In contrast, *Average-prediction* averages the predictions from individual frames. The *Max-prediction* method selects the single-frame prediction with the highest confidence as the final decision. And the *Vote* technique employs a voting mechanism across all frame predictions to determine the final result. Table 3 presents the results of these integration strategies. Our proposed method outperforms the alternatives, as the attention mechanism assigns varying weights to each frame, effectively reducing ambiguity in predictions while retaining critical information.

Moreover, we modify the input to our policy module during training to CLIP features, emphasizing the significance of disentangling semantic features for active open-vocabulary recognition. This alteration, however, adversely affects performance across all class splits, particularly in novel classes. The underlying reasons are twofold. First, while CLIP features offer a robust representation for classification, they may not be ideally suited for determining subsequent movements. Second, a policy input with strong semantic knowledge can potentially hinder the agent's performance in unfamiliar categories during testing.

Table 3. Ablation studies of evidence integration strategies and the policy input, evaluating their success rates at the final step.

Method	Base classes		Novel classes		Open classes	
	top-1	*top-3*	*top-1*	*top-3*	*top-1*	*top-3*
Ours	**60.6**	**81.3**	**36.6**	**55.1**	**53.3**	**73.4**
w/ *Average-feature*	57.5	78.1	34.7	54.0	50.6	70.9
w/ *Average-prediction*	58.9	80.0	35.9	54.3	51.9	72.2
w/ *Max-prediction*	58.6	79.9	35.2	53.7	51.6	72.0
w/ *Vote*	58.8	80.1	35.3	55.2	51.7	72.8
Train w/ CLIP feature	52.6	74.7	28.6	46.0	45.4	66.1

6. Limitation and Future Work

In our experiments, the principal challenge we addressed is the agent's ability to handle varying viewpoints. However, real-world recognition scenarios often present a broader and more complex array of challenges, such as varying lighting conditions. We leave the investigation for our future work.

7. Conclusion

This paper is driven by dual motivations: firstly, to enhance the capabilities of active recognition agents in handling open vocabulary using recent CLIP models, and secondly, to overcome the inherent limitations of CLIP in unconstrained embodied perception scenarios. We begin by evaluating CLIP models against various viewpoints and occlusions, noting their sensitivity to sub-optimal viewing conditions. To mitigate this, we present an active open-vocabulary recognition agent that intelligently acquires visual data, incorporating a self-attention module to prioritize key frames and maintain critical information globally. By employing similarity measures, we reduce class biases, improving robustness to new classes. Our empirical and ablation studies validate the effectiveness of the proposed agent on both datasets.

Acknowledgements

This work was supported in part by National Science Foundation grant IIS-2007613.

References

[1] John Aloimonos. Purposive and qualitative active vision. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, pages 346–360. IEEE, 1990. 2

[2] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International journal of computer vision*, 1(4): 333–356, 1988.

[3] Alexander Andreopoulos and John K Tsotsos. A theory of active object localization. In *IEEE International Conference on Computer Vision*, 2009. 2

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 7

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 3, 7

[6] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020. 1, 2

[7] Xiangyu Chen, Zelin Ye, Jiankai Sun, Yuda Fan, Fang Hu, Chenxi Wang, and Cewu Lu. Transferable active grasping and real embodied dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3611–3618. IEEE, 2020. 1

[8] Ricson Cheng, Ziyan Wang, and Katerina Fragkiadaki. Geometry-aware recurrent neural networks for active visual recognition. *arXiv preprint arXiv:1811.01292*, 2018. 1, 2

[9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 2, 3

[10] Wenhao Ding, Nathalie Majcherczyk, Mohit Deshpande, Xuewei Qi, Ding Zhao, Rajasimman Madhivanan, and Arnie Sen. Learning to view: Decision transformers for active object detection. *arXiv preprint arXiv:2301.09544*, 2023. 1, 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[12] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3583–3592, 2016. 2

[13] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Is first person vision challenging for object tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2698–2710, 2021. 3

[14] Lei Fan and Ying Wu. Avoiding lingering in learning active recognition by adversarial disturbance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4612–4621, 2023. 2

[15] Lei Fan, Peixi Xiong, Wei Wei, and Ying Wu. Flar: A unified prototype framework for few-sample lifelong active recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15394–15403, 2021. 1, 3, 5

[16] Zhaoyuan Fang, Ayush Jain, Gabriel Sarch, Adam W Harley, and Katerina Fragkiadaki. Move to see better: Self-improving embodied object detection. *arXiv preprint arXiv:2012.00057*, 2020. 2

[17] Samir Yitzhak Gadre, Kiana Ehsani, Shuran Song, and Roozbeh Mottaghi. Continuous scene representations for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14849–14859, 2022. 2

[18] Dimitrios Gallos and Frank Ferrie. Active vision in the era of convolutional neural networks. In *2019 16th Conference on Computer and Robot Vision (CRV)*, pages 81–88. IEEE, 2019. 2

[19] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[21] Dinesh Jayaraman and Kristen Grauman. Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion. In *European Conference on Computer Vision*, 2016. 1, 2, 3, 5

[22] Dinesh Jayaraman and Kristen Grauman. End-to-end policy learning for active visual categorization. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1601–1614, 2018. 2, 3

[23] Edward Johns, Stefan Leutenegger, and Andrew J Davison. Pairwise decomposition of image sequences for active multi-view recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3813–3822, 2016. 2

[24] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838, 2022. 1, 3

[25] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 1

[26] Klemen Kotar and Roozbeh Mottaghi. Interactron: Embodied adaptive object detection. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 14860–14869, 2022. 2

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 1

[28] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems*, 35:32340–32352, 2022. 1, 3

[29] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 2

[30] David Nilsson, Aleksis Pirinen, Erik Gärtner, and Cristian Sminchisescu. Embodied visual active learning for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2373–2383, 2021. 2

[31] Anwesan Pal, Yiding Qiu, and Henrik Christensen. Learning hierarchical relationships for object-goal navigation. In *Conference on Robot Learning*, pages 517–528. PMLR, 2021. 1

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5, 7

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 3

[34] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 2

[35] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2, 3

[36] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 6

[37] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, pages 492–504. PMLR, 2023. 3

[38] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 3

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[40] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34:251–266, 2021. 1, 2, 3, 4, 7

[41] Mitchell Wortsman, Kiana Ehsani, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[42] Zhirong Wu, Shuran Song, Aditya Khosla, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets for 2.5 d object recognition and next-best-view prediction. *arXiv preprint arXiv:1406.5670*, 2 (4), 2014. 2

[43] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 2, 3

[44] Jianwei Yang, Zhile Ren, Mingze Xu, Xinlei Chen, David J Crandall, Devi Parikh, and Dhruv Batra. Embodied amodal recognition: Learning to move to perceive objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2040–2050, 2019. 1, 2

[45] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 1

[46] Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision*, 130(10):2425–2452, 2022. 3

[47] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022. 2

[48] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2

[49] Qianfan Zhao, Lu Zhang, Bin He, Hong Qiao, and Zhiyong Liu. Zero-shot object goal visual navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2025–2031. IEEE, 2023. 1, 3