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

Supplementary Material

Abstract

This document serves as supplementary material for "Active Open-Vocabulary Recognition: Let Intelligent Moving Mitigate CLIP Limitations". It begins with a detailed examination of the dataset collected to check the limitations of CLIP models. This is followed by an extensive presentation of investigation results for various CLIP models, which could not be fully included in the main paper due to the page constraint. The material then outlines the training procedures and hyperparameters employed during our training. Lastly, it provides a thorough quantitative and qualitative analysis of the results, alongside relevant statistics, underscoring the efficacy of our proposed method.

1. Investigation Dataset

ShapeNet dataset. We select the training split of the ShapeNetCore dataset [1] for our investigation. This split contains approximately 41500 CAD models, with the class distribution illustrated in Figure 1.

For each object, a 12×12 viewing grid is uniformly sampled at intervals of 30 degrees. We set the resolution of each view to 320×320 pixels. In testing various CLIP models [2, 6], we utilize their respective image preprocessing methods prior to input into the model.

Habitat dataset. We select objects from 25 different classes within 145 semantically annotated scenes from the Habitat HM3D dataset [4], totaling 4659 objects. The distribution of these selected object categories is illustrated in Figure 2.

For each target object, the agent is initially positioned at a random location within a 3-meter radius of the target. Subsequently, the agent is rotated in 30-degree increments around the target, maintaining a constant distance, on the same horizontal plane. This procedure allows for a maximum of 12 unique viewpoints for each object. However, due to the natural occurrence of occlusions in indoor environments, viewpoints where the target is not visible are discarded.

2. Result of CLIP on Varying Viewpoints

This section extends the results presented in Section 3.2 of our main paper, providing additional insights. Specifically, we analyze the accuracy of different viewpoints using the collected ShapeNet dataset, as depicted in Figures 3 and 4. For each class, we present one example testing sample accompanied by a heatmap that illustrates the accuracy for each viewpoint. It is important to note that the accuracy for each view represents the average accuracy across all testing samples. These results are obtained using the ViT-B/32 architecture.

A key observation from our study is that the performance of CLIP is notably influenced by changes in viewpoints. In other words, certain specific viewpoints are more appropriate for object recognition compared to other. This phenomenon is in line with human perception, wherein we tend to recognize objects from their most distinctive views. In the context of embodied perception, the environment is often highly unconstrained, and the setup for capturing images by an agent is typically not within human control. This means the initial viewpoint for on-agent recognition modules might be undesired. Our findings suggest that when deploying CLIP models in embodied agents, there is a critical need to actively seek novel and informative observations, rather than relying passively on a single visual input.

Additionally, we present a comparative analysis of the mean, median, and maximum accuracy across different CLIP architectures and viewpoints in Figure 5. Specifically, we analyze four distinct CLIP architectures: two based on Vision Transformers (ViT-B/32, ViT-L/14), one ResNet-50 model (RN50x64) [2], and the recent MetaCLIP model [6]. This analysis reveals variations in performance among different CLIP models. Moreover, it is evident that all examined CLIP models demonstrate a preference for specific viewpoints, which underscores the significance of viewpoint selection in embodied AI systems. We also provide the result of testing four models on the collected Habitat dataset, which is shown in Figure 6.

3. Result of CLIP on Occlusions

In this section, we extend the results presented in Section 3.3 of our main paper. We conduct a comprehensive study of various CLIP models, namely, ViT-B/32, ViT-L/14, RN50x64 [2], and MetaCLIP [6], focusing on their resilience to randomly occurring occlusions. Specifically, we examine occlusion levels set at 20%, 35%, and 50%. The findings are illustrated in Figure 7.

We observe a noticeable degradation in the performance of CLIP models as the level of occlusion increases. Given that occlusions are common in embodied perception scenarios, it appears necessary to mitigate them by actively altering viewpoints, particularly when deploying CLIP models in embodied agents.

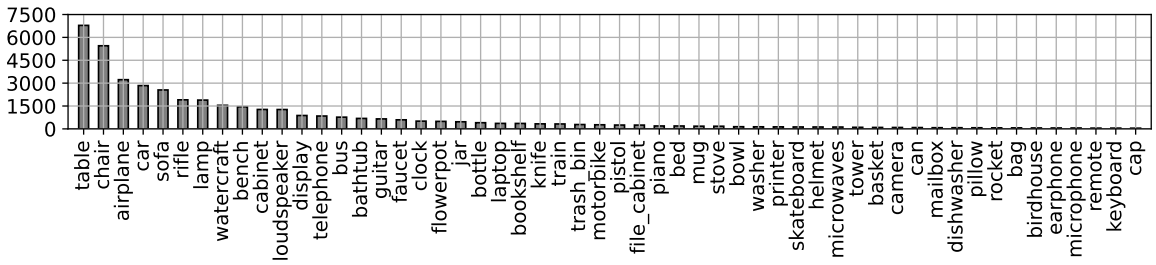


Figure 1. Instances for each category in the ShapeNet dataset for investigation.

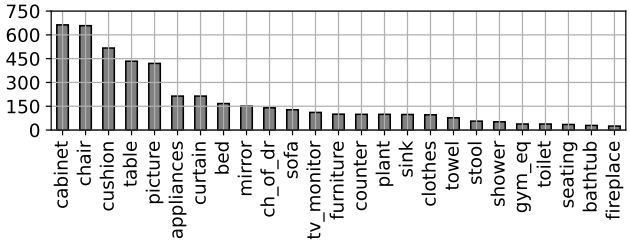


Figure 2. Instances for each category in the collected Habitat dataset for investigation. The ch_of_dr and gym_eq are short for "chest of drawers" and "gym equipment", respectively.

4. Implementation Details

In this section, we discuss the implementation details of our proposed active open-vocabulary recognition agent.

Inputs. For CLIP models, they require the computation of cosine similarity between text and visual embeddings. Consequently, descriptive texts for potential categories are indispensable at both the training and testing stages. We utilize a format comprising 'article + noun' for text descriptions of each class, such as 'a table' or 'a chair'.

Regarding visual inputs, the original resolution of the ShapeNet dataset is set at 320×320, in contrast to 800×640 for the Habitat dataset. For the Habitat dataset, recognition focus solely on the cropped window of the target, based on the presumption that the agent has pre-existing knowledge of the target's 2D position in the current frame and is tasked with classification. In our Habitat dataset experiments, the ground-truth tracking box is directly employed. For real-world applications, the agent may utilize a class-agnostic visual tracker [5], or employ a depth sensor to map the target from the initial frame to the current frame.

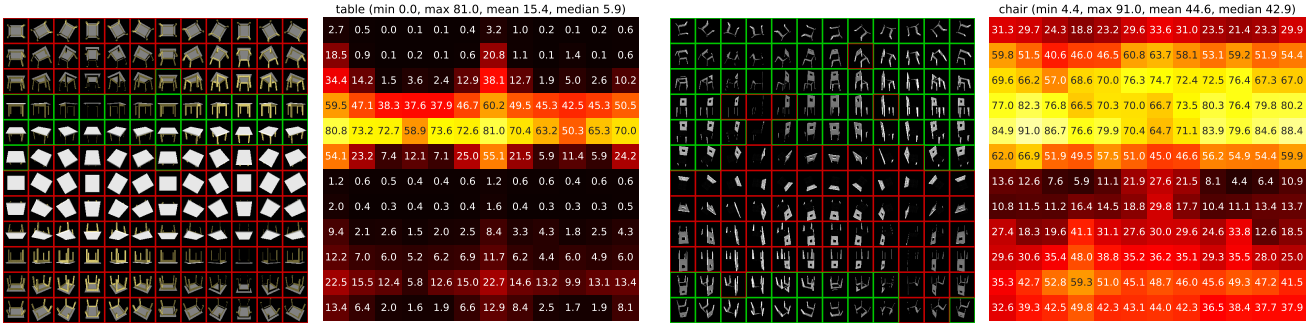
Additionally, for the separate policy image encoder, visual inputs are resized to 224×224 for both datasets.

Architecture. We primarily focus on detailing the proposed evidence integration and policy components. The evidence integration component utilizes a self-attention module, which assigns weights to frames based on their importance. Specifically, the self-attention module receives one-dimensional features, q_t, derived from concatenating three key elements: inter-concept similarity s_t^{concept}, inter-frame similarity s_t^{frame}, and proprioceptive knowledge p_t. For s_t^{concept}, we compute the top-k similarity measure between the image feature embeddings and the corresponding text embeddings. Throughout our experiments, we set the value of k to 10. The final integrated feature for recognition is then calculated as a weighted sum of all collected CLIP features, with the weights determined by the module.

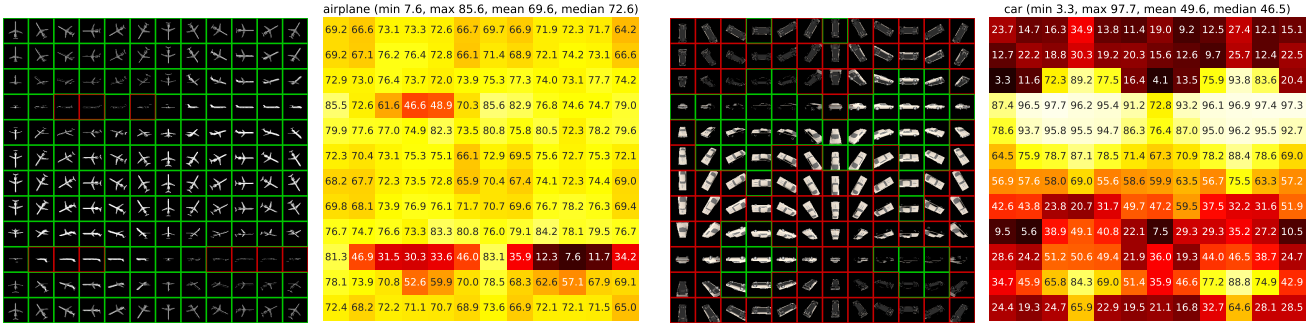
Regarding the policy component, it features a straightforward, three-layer convolutional network as the image encoder. The policy employs a single-layer GRU for integrating temporal information. This is further complemented by two distinct linear layers, serving as the actor and critic respectively.

Training. Our proposed agent comprises two trainable components: the evidence integration and the policy modules. During training, these modules are jointly optimized. This approach is chosen as the reward for the policy part is derived on predictions from individual frames rather than on integrated predictions. This strategy is free from any adverse interactions between the two modules in the initial stages of training. In our experiments, the batch sizes are set to 16 for the ShapeNet dataset and 30 for the Habitat dataset.

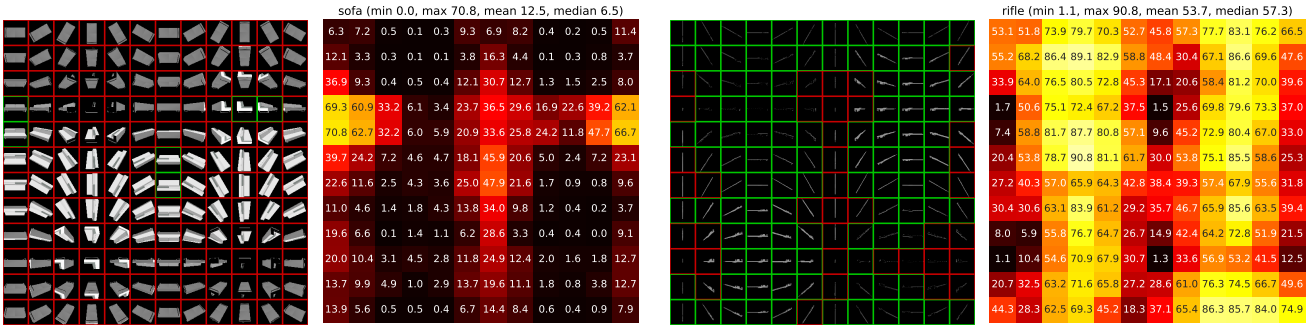
For optimizing the integration part, we utilize Stochastic Gradient Descent (SGD) with a learning rate of 10^{-5}, a momentum factor of 0.9, and a weight decay parameter set to 0.0005. Conversely, the policy module is optimized using the Proximal Policy Optimization (PPO) algorithm [3].



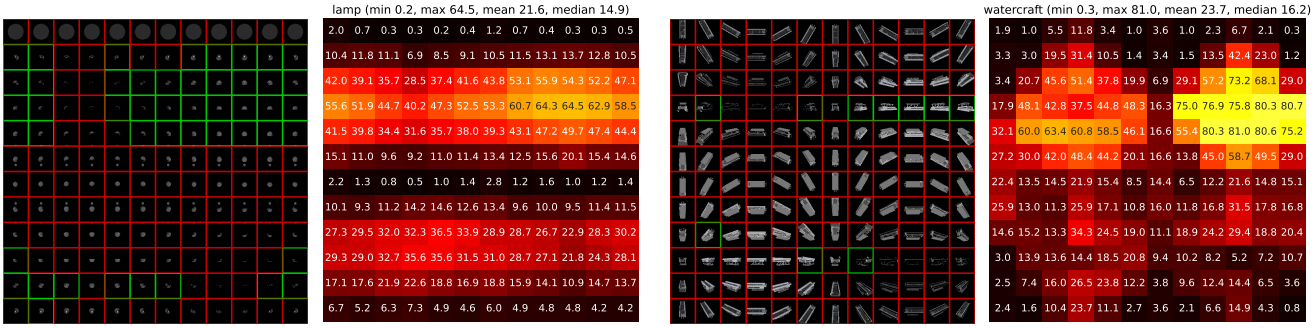
(a) Table (b) Chair



(c) Airplane (d) Car



(e) Sofa (f) Rifle

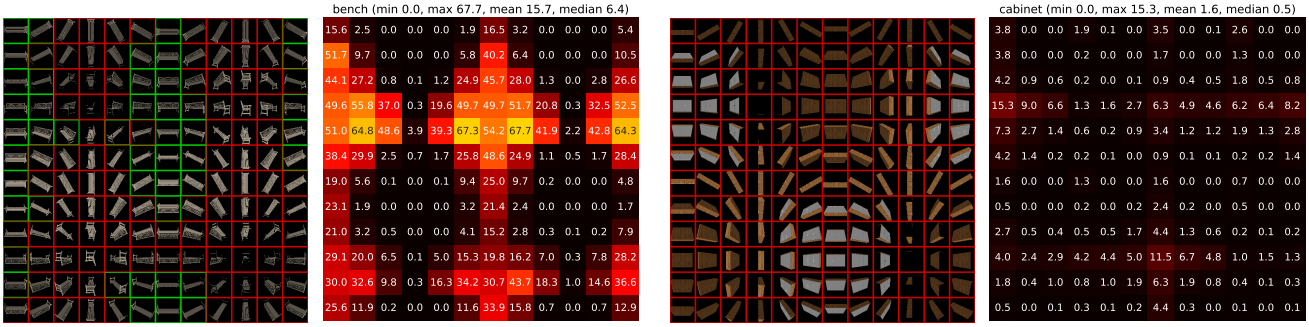


(g) Lamp (h) Watercraft

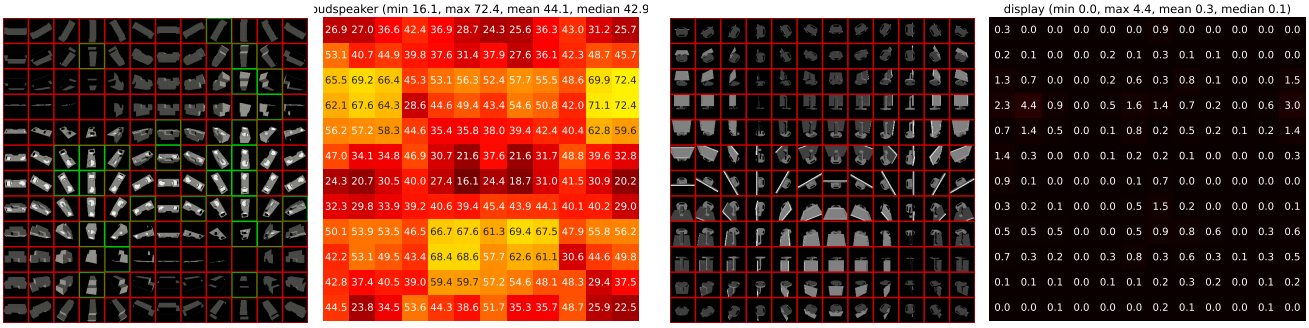
Figure 3. Recognition accuracy of different viewpoints on the most common sixteen categories (part 1).

Here, the Adam optimizer is employed with a learning rate of 2.5×10^{-5} and an epsilon value of 5×10^{-5}. Additionally, we set the discount factor γ to 0.99 for computing returns.

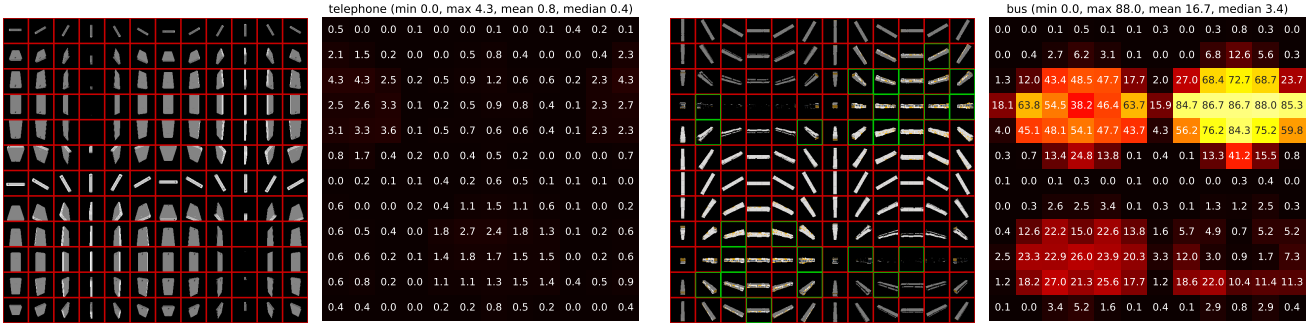
Moreover, the training for the proposed agent to achieve convergence are approximately 24 hours for the ShapeNet dataset and 30 hours for the Habitat dataset, respectively.



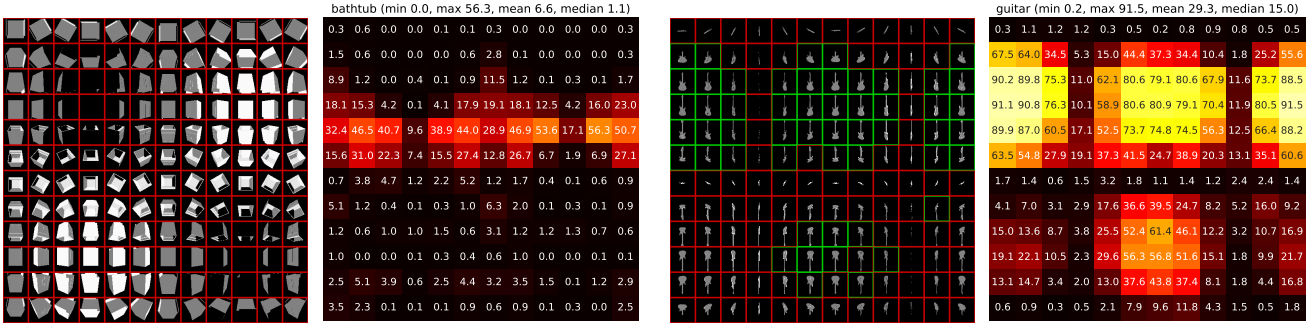
(a) Bench (b) Cabinet



(c) Loudspeaker (d) Display



(e) Telephone (f) Bus



(g) Bathtub (h) Guitar

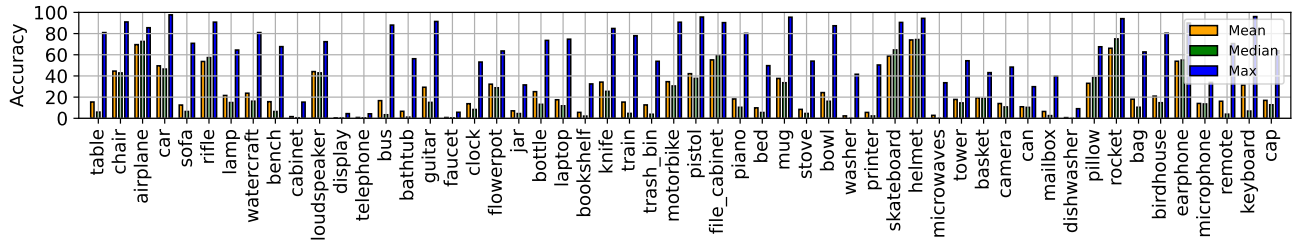
Figure 4. Recognition accuracy of different viewpoints on the most common sixteen categories (part 2).

5. Quantitative Results and Analysis

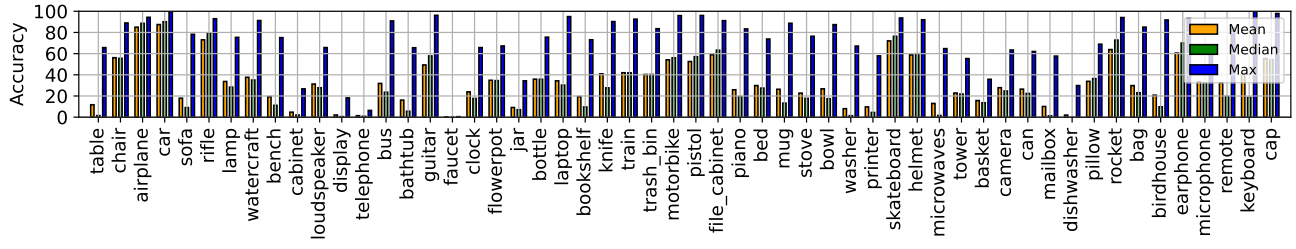
In this section, we provides more experimental analysis of the proposed active open-vocabulary recognition method.

5.1. Categorical Accuracy

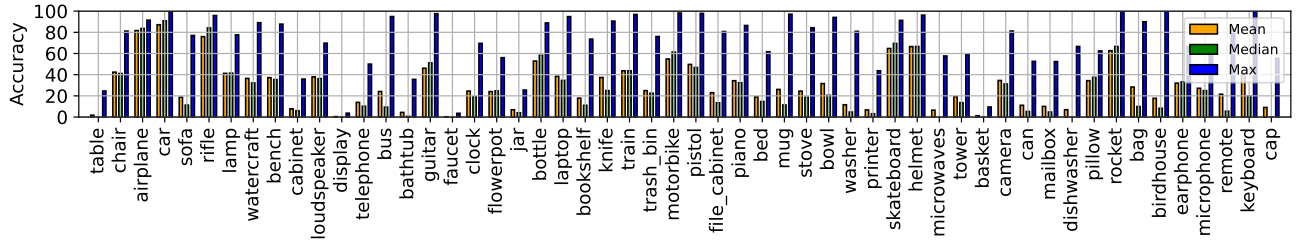
Figure 8 illustrates the recognition accuracy across various object categories within the ShapeNet dataset. Results are reported for two distinct class splits, namely, the $10/45/55$



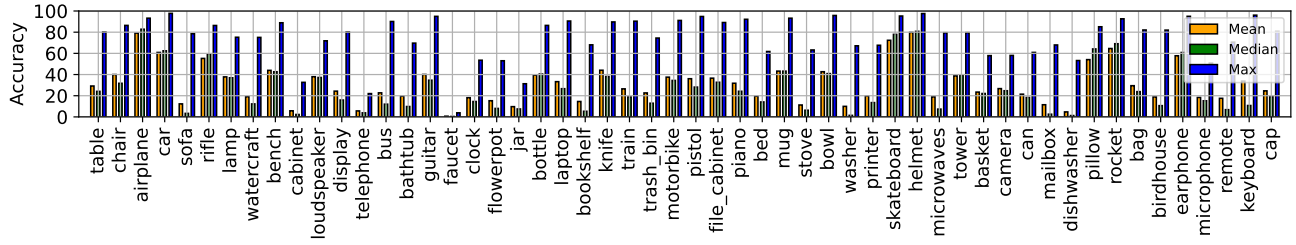
(a) ViT-B/32 [2]



(b) ViT-L/14 [2]



(c) RN50x64 [2]



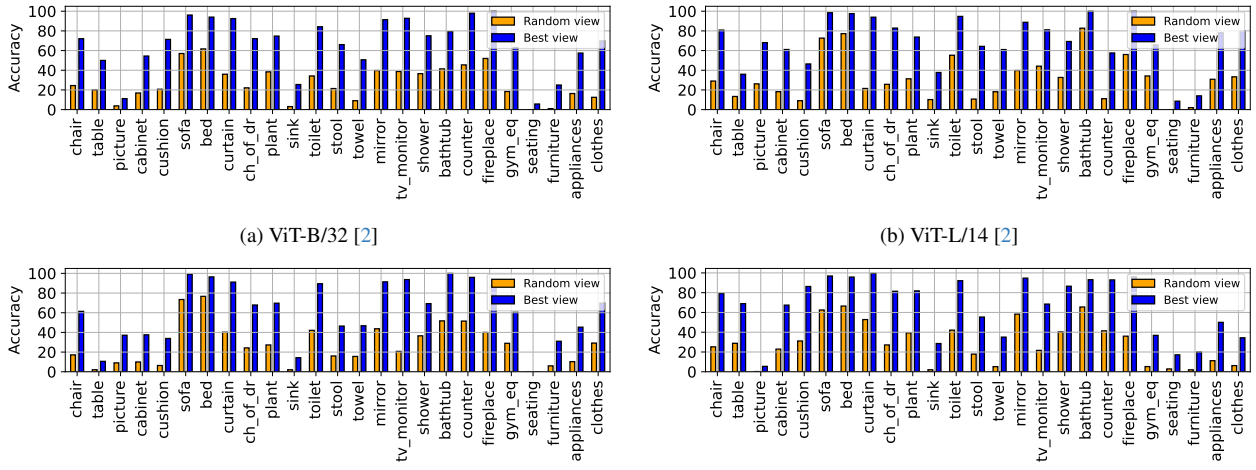
(d) MetaCLIP [6]

Figure 5. Performance of various CLIP models on the collected ShapeNet dataset: reporting mean, median, and maximum accuracy across all viewpoints for each category.

and the $20/35/55$ splits. As the proposed agent engages in active perception of the environment, an improvement in recognition accuracy is observed across all classes. This underscores the efficacy of the proposed method in effectively managing both base and novel unseen classes.

5.2. View Visiting Frequency

We provide two examples of view visiting frequencies in Figure 9. To calculate the view visiting frequency, we record the location of our agent at each steps during recognizing samples belonging to a specific category. Intuitively, a intelligent movement policy would try to reach views that the equipped CLIP model could correctly classify the target. In other words, the visiting frequencies for views that could be recognized should be more frequently visited. The visiting frequency is shown on the right of each example in Figure 9, and we normalize the frequency to $[0, 100]$ for each class. It is worth noting that, for each recognition episode, the starting viewpoint is randomly sampled among all 12×12 viewpoints, meaning the agent needs to execute different actions to reach informative viewpoints.



(a) ViT-B/32 [2]

(b) ViT-L/14 [2]

(c) RN50x64 [2]

(d) MetaCLIP [6]

Figure 6. Performance of CLIP models on the collected Habitat dataset. This table illustrates changes in performance based on different viewpoints. The comparison is conducted on a per-object basis, *i.e.*, comparing the accuracy of a randomly chosen viewpoint against that of the best-performing viewpoint. The average performance across all test samples within each category is then calculated.

5.3. Ablation Study on q_t

We test our integration module by selectively masking components in q_t. The experiment is conducted using the ShapeNet with 10 base classes. To enhance the comparability, we introduce the hit rate metric. The hit rate metric computes the percentage of episodes where the highest α_t corresponds to a correct prediction. A higher hit rate indicates more effective α prediction. We observe that our integration module exhibits higher sensitivity to s_t^{frame}, which measures the temporal similarity between the current frame and all preceding frames. Note that the behavior of the policy remains consistent.

Table 1. The ablation study on the proposed integration module by selectively masking components in q_t.

	Base classes		**Novel classes**		**Open classes**		**Hit rate**
Ours	**60.6**	**81.3**	**36.6**	**55.1**	**53.3**	**73.4**	**65.5**
w/o s_t^{concept}	59.3	80.9	35.5	55.0	52.1	73.0	60.2
w/o s_t^{frame}	58.9	80.1	34.3	53.8	51.2	72.2	58.2
w/o p_t	59.5	81.0	35.7	55.0	52.3	73.0	62.2

6. Qualitative Results

We present additional qualitative results of the proposed agent, demonstrated through a video encompassing both datasets. Initially, we delve into the limitations inherent to CLIP models, particularly in the context of embodied perception scenarios. Subsequent to this, we showcase testing episodes of our proposed methodology. During each testing episode on the ShapeNet dataset, the agent's current location, next movement, and its predictions are illustrated. For comparative analysis, baseline CLIP predictions are also provided. This comparison allows for an observation of the enhancements in our proposed active open-vocabulary recognition agent, as it progresses through successive steps. The demonstrative results on the Habitat dataset are also included.

References

[1] 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. 1

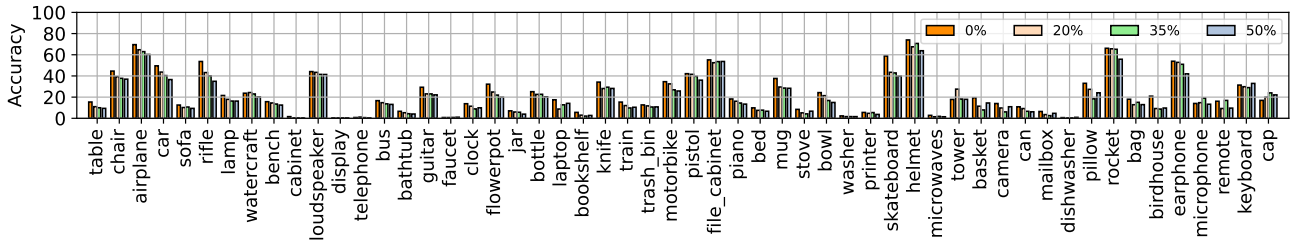
[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 5, 6, 7

[3] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

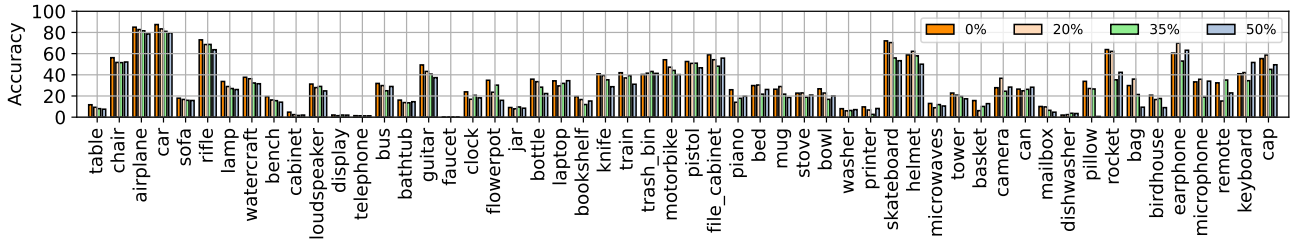
[4] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34:251–266,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

[5] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019. 2

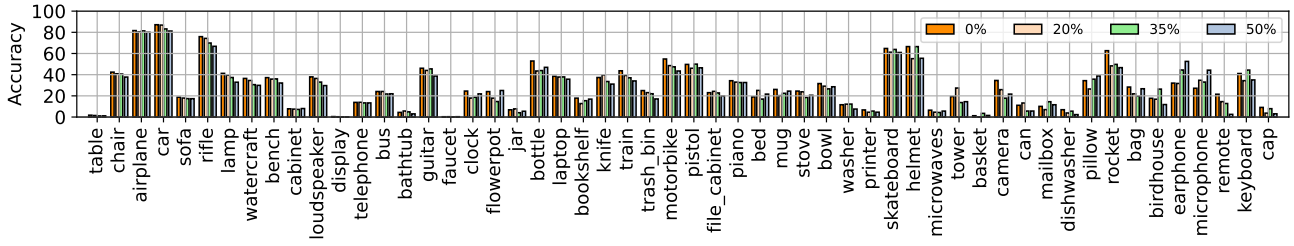
[6] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Rus-



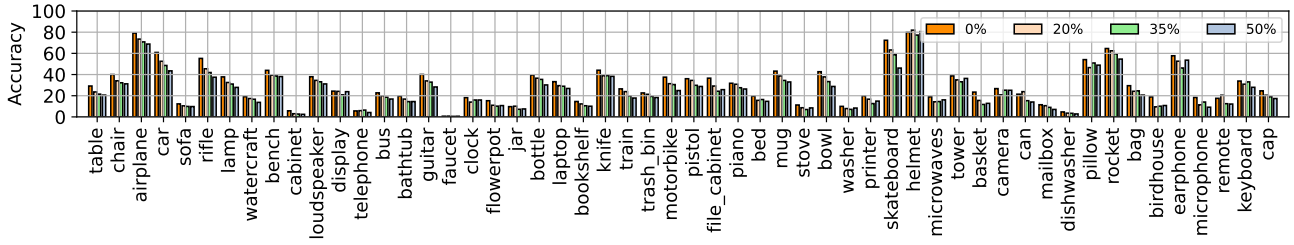
(a) ViT-B/32 [2]



(b) ViT-L/14 [2]



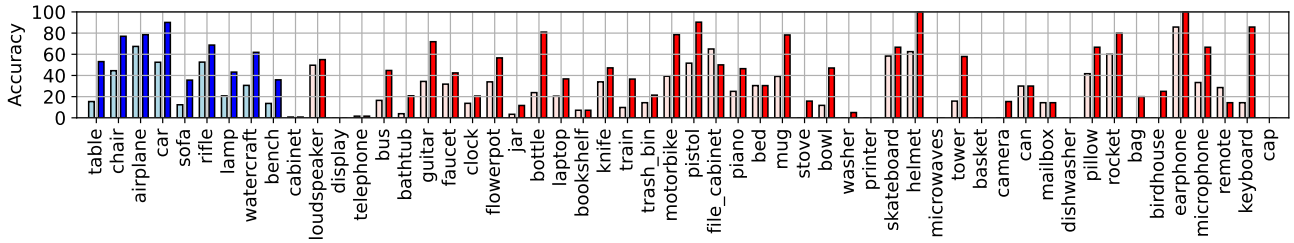
(c) RN50x64 [2]



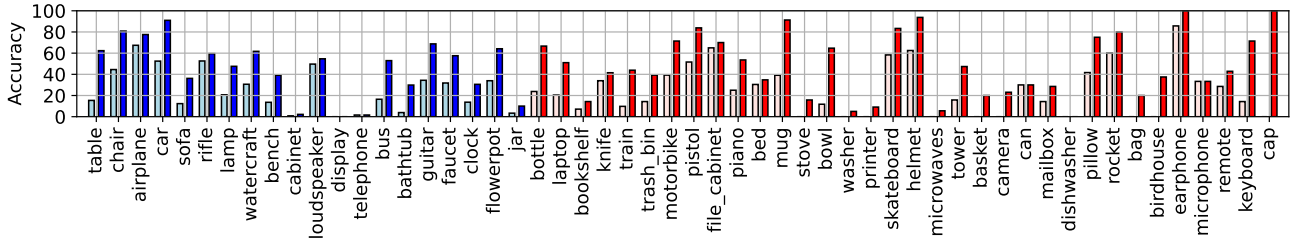
(d) MetaCLIP [6]

Figure 7. Performance of various CLIP models on the adversarial inference from different levels of occlusion (20%, 35%, 50%). The comparison is conducted using the collected ShapeNet dataset. Generally, mean accuracy decreases as the level of occlusion increases.

sell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 1, 5, 6, 7

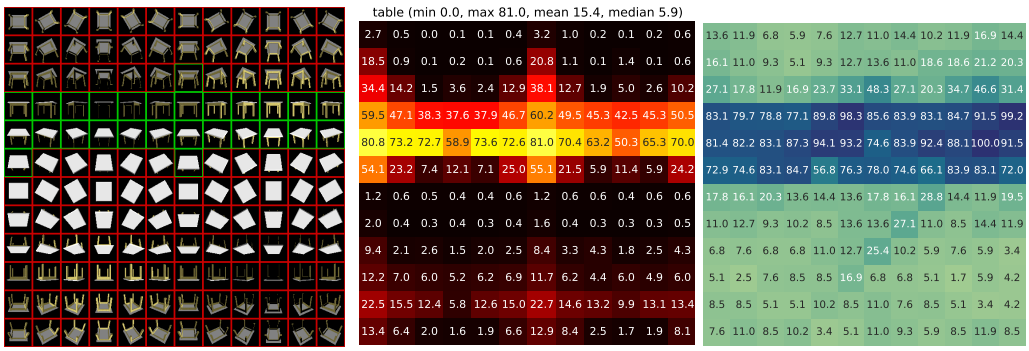


(a) With 10 base classes.

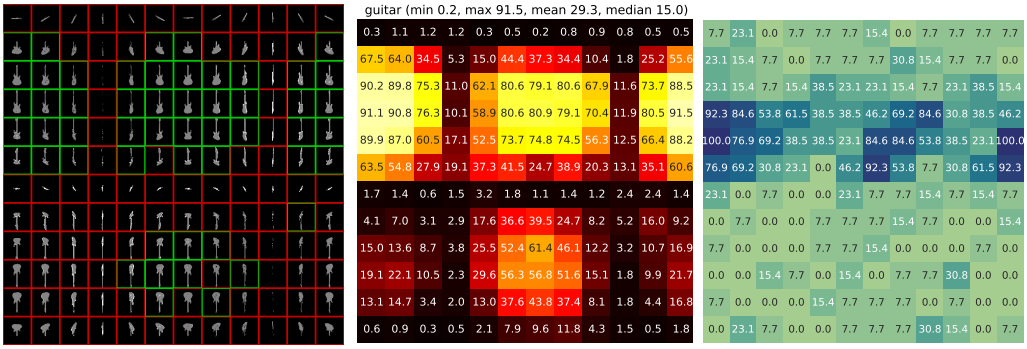


(b) With 20 base classes.

Figure 8. The recognition accuracy of each object category on the ShapeNet dataset. Base classes are denoted in blue, while novel classes are marked in red. For each class, the table presents a side-by-side comparison of the accuracy at the initial step, *i.e.*, the baseline CLIP performance (light color), and the accuracy achieved by our agent at the final step (standard color).



(a) Table



(b) Guitar

Figure 9. Visiting frequency of our agent across two example classes during the testing phase. This frequency is depicted as a heatmap, positioned to the right of each sub-figure. During testing, the starting position is randomly sampled among all viewpoints. In other words, the difference of visiting frequencies is brought by the intelligent movements. It is important to note that the visiting frequencies are normalized on a class-wise basis to facilitate a more distinct comparison. Additionally, a testing sample is included alongside the baseline performance obtained from CLIP, serving as the reference.