Reasoning Mamba: Hypergraph-Guided Region Relation Calculating for Weakly Supervised Affordance Grounding

Supplementary Material

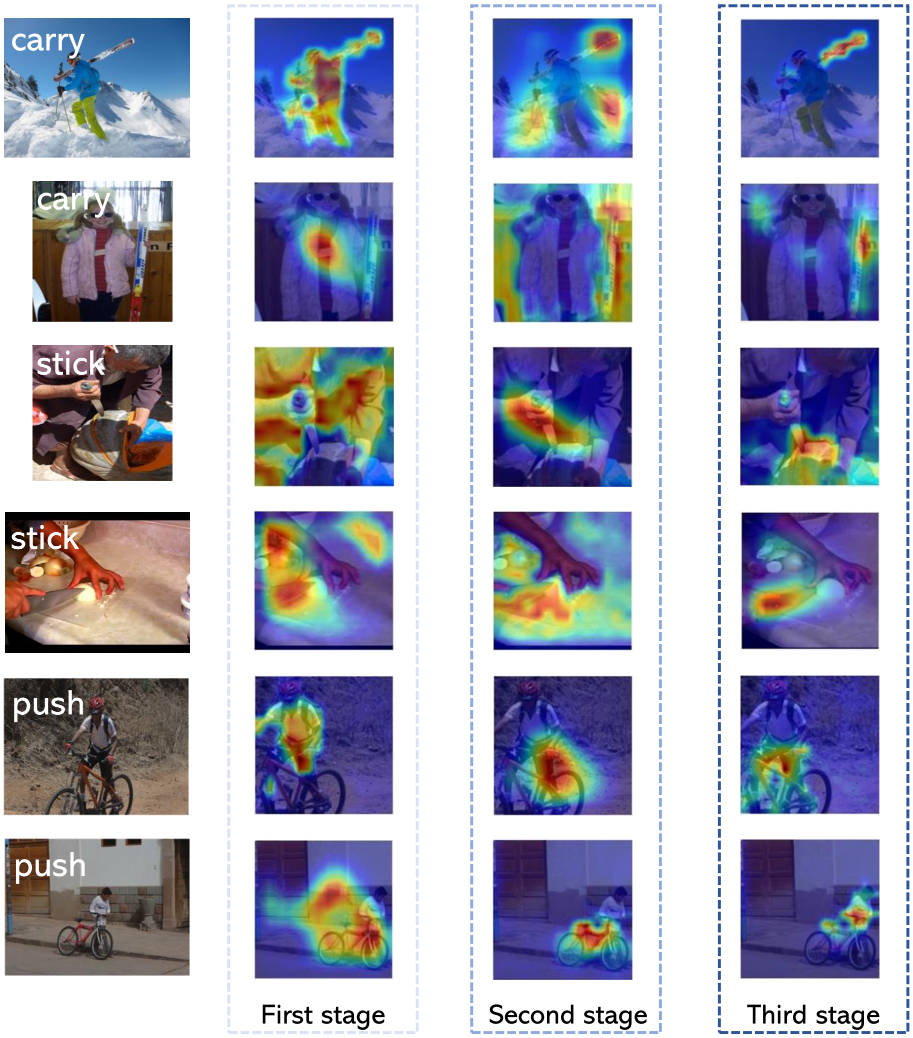

Figure 1. Visualization of the hypergraph evolution process. The results demonstrate that the model progressively focuses on the object-related affordance regions.

1. Analysis for Hypergraph Evolution

The Hypergraph Evolution Module is designed to optimize the existing hypergraph structure to capture high-order relationships within the data more effectively. To achieve this goal, we propose a three-stage hypergraph evolution process, where each stage contributes to improving the adaptability and expressiveness of the hypergraph.

In the first stage, we dynamically introduce new hyperedges based on the semantic space. Specifically, we analyze the semantic features of the input image to identify potential associations and incorporate new hyperedges to complement the existing hypergraph structure. This process enhances the hypergraph's representational capacity, better capturing the complex relationships among regions. In the second stage, we utilize egocentric deep features as a reference to filter the most relevant vertices and hyperedges associated with the object's affordance cues. This approach

Method	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Clustering	1.212	0.403	1.192	1.398	0.375	1.169
Hypergraph	**1.173**	**0.414**	**1.247**	**1.372**	**0.380**	**1.190**
ViT	1.196	0.408	1.221	1.380	0.377	1.180
Mamba	**1.173**	**0.414**	**1.247**	**1.372**	**0.380**	**1.190**

Table 1. Ablation results of the baselines.

Method	Inference Time	Flops	Params
R-Mamba (Ours)	0.2s	280G	32M
WSMA	0.1s	238G	51M
WSMA+Ours	0.26s	490G	82M

Table 2. Comparsion on efficiency (on batch sample) among WSMA and Ours.

is based on the observation that affordance features of different objects exhibit varying degrees of saliency in visual perception. Therefore, when constructing the hypergraph, it is essential to prioritize hyperedges directly related to the target object. This selection mechanism ensures that the hypergraph structure aligns with the requirements of the target task while effectively reducing interference from irrelevant information. In the third stage, we refine the hypergraph structure by eliminating redundant or less informative hyperedges introduced in the previous stages. To achieve this, we apply specific criteria to identify and remove hyperedges that contribute minimally to the final task. As a result, the refined hypergraph remains compact while effectively preserving key relationships within the image.

To validate the effectiveness of the hypergraph evolution process, we conduct a series of experiments by feeding features extracted from different evolution stages into subsequent modules and visualizing the corresponding heatmaps in Figure 1. If the heatmaps generated from hypergraph-computed features demonstrate a stronger focus on affordance regions, the hypergraph effectively integrates features from different regions and enhances the model's understanding of structural and relational information within the image. These experimental results further confirm the effectiveness of our proposed hypergraph evolution module in focusing on affordance-relevant visual components.

2. Implementation Details

For HICO-IIF dataset, we use SGD with a learning rate of 1e-3, weight decay of 5e-4, and a batch size of 16. The loss weight coefficients, λ_{sim} and λ_{gc}, are set to 1 and 0.15, respectively. The nearest S number is set to 1. For the WSMA [3] method, we follow its basic configuration and apply our approach to the DINO-ViT-S backbone network accordingly.

3. Ablation Study of Baselines

To further demonstrate the effectiveness of our approach, we conduct a comparative analysis between hypergraph and conventional clustering [2], as well as Mamba and ViT [1]. The results presented in Table 1 indicate that the hypergraph and Mamba architectures exhibit superior performance. This is because affordance regions are not isolated but involve region-to-region relationships with other object components. For example, the affordance of a cup is not solely determined by its body or handle in isolation but by their combination. The presence of the handle enables a stable grip, while the cup body holds liquid, and together, they define the cup's "pouring" affordance. In our approach, the DINO-ViT-S backbone segments both the handle and the cup's body into multiple patches, where their relationships are established through interactions among these patches to enable the pouring action collectively. Therefore, to accurately localize affordance regions, it is crucial to capture the many-to-many mapping relationships between different object components. While the K-means method [2] primarily focuses on grouping data, hypergraphs, in contrast, establish many-to-many mappings between regional features by connecting multiple vertices through hyperedges. This enables a more comprehensive representation of the relationships among object components. Moreover, since ViT focuses on all tokens, it struggles to effectively capture the region relationships between visual components constructed in the hypergraph. In contrast, Mamba excels at efficiently modeling sequence dependencies, making it well-suited for processing hypergraph-based information. It captures the connections between local structures and global semantics, which is critical for affordance understanding. Therefore, we adopt the Mamba architecture to implement this process.

4. Results Analysis on WSMA

Through an in-depth analysis of the efficiency comparison experiments, our method in Table 2 demonstrates significant advantages in model complexity control. Compared to the WSMA method, our method successfully reduces the number of params by approximately 37%. Notably, despite the substantial reduction in parameter size, our method incurs an increase in inference time compared to the baseline method. This phenomenon arises from the high-order interaction paradigm of hypergraph computation. Unlike traditional pairwise point interactions, hyperedge operations update the joint states of multiple vertices, which enhances the model's representational capacity but inevitably increases the computational steps in each forward pass.

References

[1] 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.

[2] Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[3] Lingjing Xu, Yang Gao, Wenfeng Song, and Aimin Hao. Weakly supervised multimodal affordance grounding for egocentric images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6324–6332, 2024.