LoCo: Selective Local Competition for Discriminative Open-Vocabulary Multi-Label Recognition

Zhuoming Li[1] **Bo Han**[2] **Xiaoyu Wang**[1] **Yuheng Jia**[2,3]*
[1]School of Software Engineering, Southeast University, Nanjing, China
[2]School of Computer Science and Engineering, Southeast University, Nanjing, China
[3]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
{leezhuoming,hanbo,wang_xy,yhjia}@seu.edu.cn

Abstract

Open-Vocabulary Multi-Label Recognition (OV-MLR) aims to identify all queried semantic concepts present in an image. Existing methods mainly improve global image-text matching or text-side representations, leaving local prediction largely under-explored. In this paper, we revisit local prediction and reveal that it can serve as a strong discriminative branch once two properties are properly handled: spatial consistency of local visual features and compatibility-aware local competition. This perspective is motivated by local semantic sparsity: each image region is related to only a small subset of queried concepts, while compatible concepts of different granularities or types may still coexist locally. We instantiate this perspective with **LoCo**, a selective **Lo**cal **Co**mpetition framework for OV-MLR. LoCo adopts a spatially consistent visual encoder to reduce local feature entanglement, and introduces Selective Softmax to impose competition only among locally incompatible categories while preserving compatible responses. Multi-scale inference further improves local prediction at different scales. LoCo can also be ensembled with an existing global predictor to provide complementary image-level context. Experiments on four multi-label benchmarks show that standalone LoCo improves average F1-score and mAP over TagCLIP by **3.4%** and **3.8%**, respectively. LoCo + DualPrompt achieves the highest averages among the compared methods, with gains of **4.8%** and **5.2%** over TagCLIP. Our results suggest that local semantic sparsity offers a useful basis for developing more discriminative open-vocabulary multi-label recognition methods. The code is available at https://github.com/lee-plus-plus/LoCo.

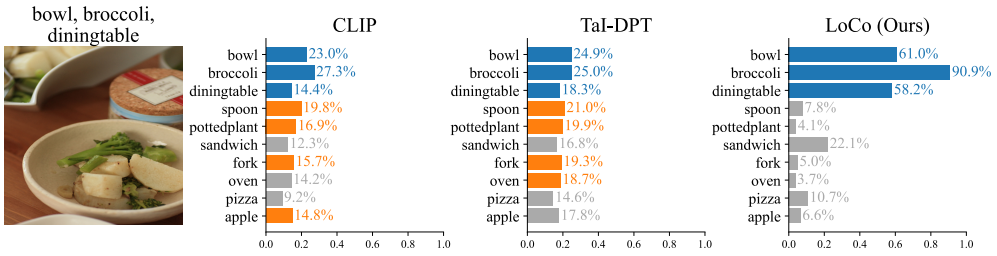
1 Introduction

Vision-language models, *e.g.*, CLIP [1], have significantly advanced open-vocabulary visual recognition by aligning images and texts in a shared semantic space, where categories can be dynamically specified by natural language. Although CLIP enables single-label open-vocabulary recognition [2–4], real-world visual understanding is inherently multi-label [5]: an image often contains multiple objects, scenes, and attributes simultaneously. Open-Vocabulary Multi-Label Recognition (OV-MLR) thus requires a model to recognize all semantic concepts present in an image, given arbitrary category names or textual descriptions, without supervised retraining [6].

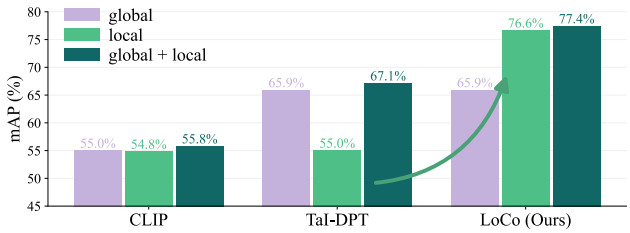
Although previous OV-MLR studies have made progress [7–14], achieving discriminative multi-label predictions remains challenging. Existing methods produce prediction scores with a relatively small margin between positive and negative categories, making it difficult to obtain discriminative predictions, as shown in Figure 1a. This phenomenon reflects two key challenges in OV-MLR: (1) **visual**

*Corresponding author.

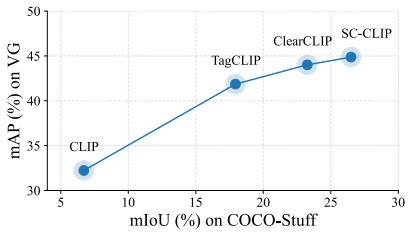
40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Existing OV-MLR methods CLIP [1] and TaI-DPT [8] produce predictions with low discriminability.



(b) Our model (LoCo) unleashes the potential of local prediction.



(c) The segmentation capability of the visual backbone enables multi-label local prediction.

Figure 1: Motivation and key insight of this paper. (a) Existing OV-MLR methods [1, 8] produce multi-label predictions with low discriminability, and our model overcomes this limitation. Blue bars denote true positive predictions, while orange bars denote false positive predictions. (b) The potential of local prediction is under-exploited by existing methods [1, 8]. Our model unleashes its potential and achieves outstanding performance. (c) The spatial consistency of the dense visual encoder, measured by its segmentation performance, *i.e.*, mIoU, enables multi-label recognition in a local prediction framework.

feature entanglement. Multiple semantic concepts are compressed into a single visual representation, which dilutes category-specific cues [15]. (2) **weak response separability**. Text features encoded by CLIP are highly correlated [16] (shown in Appendix A). Since CLIP-based category scores are computed from cosine similarities between visual features and textual queries, class-independent predictions tend to produce small numerical gaps among category scores. Existing OV-MLR methods mainly improve prediction scores from the text side [8, 17, 9, 16, 12, 13], which partially alleviates the weak response separability challenge, leaving the feature entanglement challenge relatively underexplored.

To achieve discriminative multi-label prediction, we argue that **local prediction** has the potential to address both of the aforementioned challenges. Unlike global features that include multiple coexisting semantics, local features have much sparser semantic compositions, a property we term **local semantic sparsity**, which makes local prediction different from global multi-label prediction. Previous methods [8, 17, 10, 14] usually implement local prediction as a spatial repetition of image-level prediction and use it only as a supplementary branch, leaving much of its potential under-exploited, as shown in Figure 1b. This also suggests that once the bottlenecks of visual feature entanglement and weak response separability are alleviated, local prediction can bring substantial improvements over existing state-of-the-art text-side methods.

Based on this observation, we investigate how to unleash the potential of local prediction under the local semantic sparsity prior. First, to satisfy the local semantic sparsity prior, local visual features should be spatially consistent. CLIP's [1] raw local visual features also suffer from feature entanglement, resulting in noisy spatial responses [6, 16, 18] and thus unreliable local prediction. Figure 1c provides empirical evidence that stronger spatial consistency in spatial responses (mIoU) is beneficial for local multi-label recognition (mAP). Second, to exploit local semantic sparsity, a suitable inter-class competition mechanism should be introduced. Inter-class competition can effectively enlarge the relative gap between positive and negative category prediction scores in single-label prediction, but it may incorrectly suppress compatible categories in multi-label scenarios.

To this end, we propose a selective **Lo**cal **Co**mpetition framework (**LoCo**) for OV-MLR. LoCo first leverages a spatially consistent visual encoder, SC-CLIP [19], to provide more reliable local visual features. It then designs Selective Softmax, which imposes class competition on local regions while considering both exclusiveness and compatibility between categories, thereby improving the

discriminability of multi-label scores. In addition, LoCo incorporates multi-scale inference using grid cropping to enhance the detection of small objects and local details, and can be integrated with other global prediction methods. The main contributions of this work are summarized as follows:

1. We identify spatial consistency of local visual features as a key prerequisite for effective local prediction in OV-MLR, and empirically show that spatially consistent local features can turn local prediction into a strong discriminative branch.
2. We propose Selective Softmax, a local competition mechanism that introduces competition between each category and its locally incompatible competitors. It improves response discriminability while avoiding the over-suppression of compatible concepts.
3. We develop LoCo, an open-vocabulary multi-label recognition framework that improves average F1-score and mAP over the most competitive baseline, TagCLIP [6], by 3.4% and 3.8%, respectively. LoCo also complements existing global predictors: combining it with DualPrompt's global predictions [14] further improves average F1-score and mAP by 1.4% each over standalone LoCo.

2 Related Work

Multi-label learning [5] is a fundamental yet challenging task in computer vision, which requires models to identify multiple semantic concepts present in an image. Considering the high cost of collecting multi-label annotations, previous literature has extensively focused on weakly supervised settings [20–22]. In recent years, Open-Vocabulary Multi-Label Recognition (OV-MLR) has been proposed to leverage the open-category and zero-shot capability of vision-language models [1] for multi-label learning. Early studies mainly focused on adapting CLIP from single-label recognition [3] to multi-label scenarios [7, 23]. Most existing OV-MLR methods focus on improving the quality of textual queries. For example, Text-as-Image [8, 17, 10, 9] methods convert text into proxy visual samples to obtain class prototypes that are more suitable for multi-label prediction via prompt tuning. Later works [24, 25, 12, 13, 11, 14] further exploit external language knowledge, including semantic granularity, contextual descriptions, category relations, and co-occurrence knowledge. Score calibration and compositional prompts were also discussed [26].

Apart from the aforementioned methods that mainly focus on the text side, another line of work exploits local visual information. Early OV-MLR methods [7, 8] combine global and local responses as dual-granularity prediction, which has since been widely adopted by later works [10, 9, 12, 14, 16]. These works demonstrate the value of local information, but still treat local prediction as a spatial repetition of image-level prediction. TagCLIP [6] developed a local-to-global framework that directly uses CLIP local features for open-vocabulary multi-label prediction.

A closely related area is training-free Open-Vocabulary Segmentation (OVS), which systematically studies how to obtain spatially consistent dense visual representations from CLIP [27–30]. SCLIP [18] replaces final-layer query–key attention with query–query and key–key correlations, while ClearCLIP [31] uses self-self attention and removes the final residual connection and feed-forward network (FFN). SC-CLIP [19] uses intermediate-feature similarity for self-calibration and multi-level feature fusion. Other approaches [32–34] transfer spatial information from self-supervised visual models [35] to dense vision-language prediction.

3 Analysis: What Makes Local Prediction Work?

3.1 Local Prediction Remains Auxiliary

Figure 2 illustrates a typical challenge in multi-label prediction: feature entanglement on global features weakens the discriminability of global prediction. This phenomenon motivates the widely used global-local framework in OV-MLR [7, 8]. Since local features preserve fine-grained visual evidence and avoid global aggregation, they are expected to alleviate feature entanglement and produce more discriminative category responses. However, Figure 1b shows that prediction based on CLIP's local features remains weak. Moreover, combining local predictions with global predictions brings only limited improvement, which suggests that local prediction in existing frameworks remains auxiliary, rather than serving as a reliable branch. We identify that this is mainly because local features still suffer from entanglement, as discussed below.

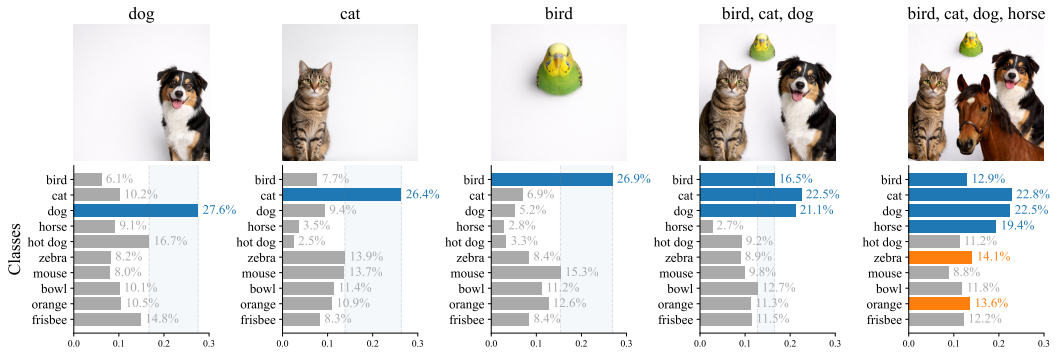
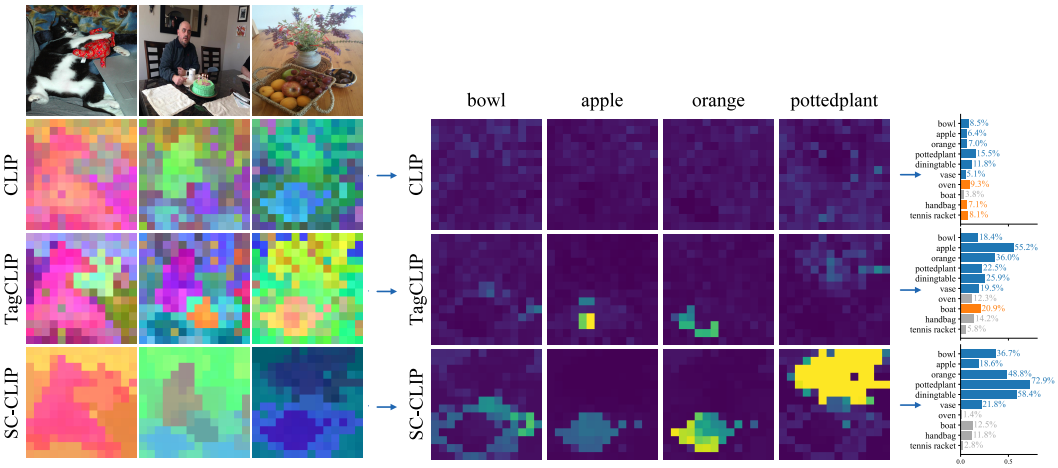


Figure 2: Feature entanglement in multi-label scenarios. Blue bars denote true positive predictions, while orange bars denote false positive predictions. As more semantic concepts co-occur in the same image, CLIP's prediction score gap between positive and negative classes narrows. We attribute this to feature entanglement caused by mutual dilution across different categories of information.



(a) t-SNE of the dense feature map. (b) Dense predictions for different classes. (c) Local scores.

Figure 3: Spatial consistency of dense visual backbones matters for discriminative local prediction. CLIP [1], TagCLIP [6], and SC-CLIP [19] are selected as weak, moderate, and strong visual backbones. Softmax is applied at each spatial position to produce dense predictions. With stronger visual backbones, local features are less fragmented and more spatially consistent, thereby producing more discriminative multi-label predictions.

3.2 Local Features Lack Spatial Consistency

Effective local prediction is based on the prerequisite that local visual features are disentangled, *i.e.*, they reflect the semantics of their corresponding image regions instead of mixing information from unrelated regions. However, this does not naturally hold for CLIP's default visual encoder, which is trained with image-text contrastive supervision without explicit spatial consistency regularization. Figure 3a shows that CLIP's local features are fragmented, spatially inconsistent, and semantically mixed. Therefore, **local features are also entangled**, which results in unreliable local activation and thus non-discriminative local prediction, as shown in Figure 3b and Figure 3c. This observation suggests that the bottleneck lies in the spatial consistency of local visual features.

To quantify the spatial consistency of local features, we adopt the semantic segmentation metric mIoU as an external proxy, which is appropriate because semantic segmentation directly evaluates whether local features can produce spatially coherent and boundary-aware category responses, matching the aforementioned requirement. Therefore, rather than designing a new spatial encoder specifically for OV-MLR, we leverage recent progress in open-vocabulary segmentation, *e.g.*, ClearCLIP [31] and SC-CLIP [19], to obtain more spatially consistent local features.

We do not claim the dense visual backbone itself as our contribution. Instead, we use it to verify an important finding: local prediction becomes effective only when local representations are spatially

Table 1: Major types of semantic relations among categories in the VG dataset.

Relation type	**Examples**	**Locally Compatible?**
Scene-level Co-occurrence	*room–bed*; *street–car*	No
Covering Relation	*person–shirt*; *hand–glove*	Usually Yes
Overlap Relation	*snow–mountain*; *light–lamp*	Usually Yes
Taxonomic Inclusion	*animal–dog*; *vehicle–car*	Yes
Part–whole Composition	*person–arm*; *car–wheel*	Yes
Surface Marking	*sign–letter*; *screen–logo*	Yes
Near-synonymy	*airplane–plane*; *cap–hat*	Yes

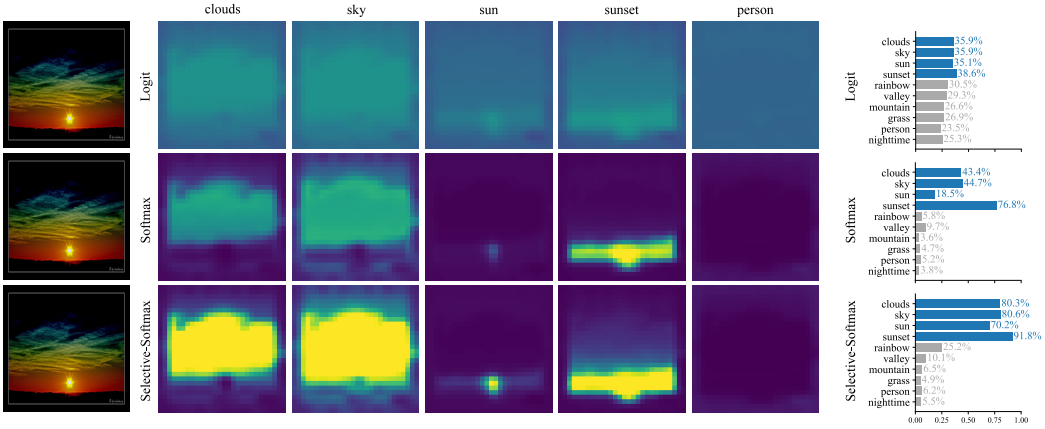


Figure 4: Multi-label local prediction with different prediction strategies. Softmax is effective for enhancing discriminability, but it introduces unreasonable competition between compatible classes, *e.g.*, *clouds* and *sky*, or *sun* and *sunset*, resulting in over-competition. Selective Softmax introduces competition between each category and its locally incompatible competitors and produces better predictions.

consistent. Figure 1c suggests a correlation between segmentation capability and MLR performance, and Figure 3 provides explainable evidence for its necessity. These facts allow us to decouple the design of dense feature representations from OV-MLR, and instead focus on how reliable local features can be effectively used for discriminative multi-label prediction in the following analysis.

3.3 From Local Over-Competition to Selective Competition

When each image region reflects its own semantics, spatially consistent local features enable local category competition in practice. Applying softmax over local category responses can enlarge relative score differences and produce more discriminative predictions, which is widely adopted in semantic segmentation. However, this strategy implicitly assumes local semantic uniqueness, *i.e.*, each local region should correspond to only one semantic category. While this assumption is natural in semantic segmentation, where each pixel is treated as a single-label prediction, it is too restrictive for OV-MLR, whose predefined category set may contain concepts of different granularities and types. For example, queried categories may include coarse and fine categories (*animal* and *dog*), or whole and part (*person* and *arm*). Table 1 summarizes the major types of semantic relationships among categories in the VG dataset, where many categories can coexist in the same region. Therefore, treating these categories as mutually exclusive would incorrectly suppress their dense predictions and limit the model's capability to handle open-vocabulary semantics. We refer to this property as **local semantic sparsity**, which assumes that each region can belong to a few concepts and is a better prior than local semantic uniqueness for OV-MLR.

To exploit local semantic sparsity, inter-class competition should be introduced selectively, *i.e.*, between each category and its locally incompatible competitors, avoiding direct competition between compatible categories. Figure 4 shows that raw logit scores are under-competitive, softmaxed scores are over-competitive, and selectively softmaxed scores provide structured local competition that better fits the semantic relationships in OV-MLR. Implementation details of Selective Softmax are described in Section 4.3.

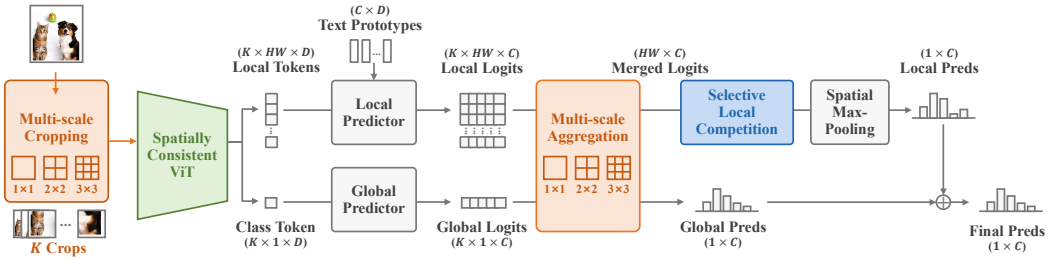


Figure 5: Framework of the proposed LoCo. An image is first divided into K crops using grid cropping and encoded by a spatially consistent ViT. Local tokens produce crop-level logits. The multi-scale aggregation module spatially aligns and averages these logits before Selective Softmax and spatial max pooling. The resulting local predictions can optionally be ensembled with global predictions averaged over the same crops.

4 Method: LoCo Framework for Multi-label Recognition

Section 3 has already shown that effective local prediction for open-vocabulary multi-label recognition requires two conditions: spatially consistent local features and a discriminative prediction mechanism that considers local exclusiveness and compatibility between categories. To achieve them, we propose **LoCo**, a selective local competition framework for open-vocabulary multi-label recognition, as shown in Figure 5. It adopts a **spatially consistent dense visual encoder** (Section 4.2) to produce local tokens for image-level local prediction. **Selective local competition** (Section 4.3) enhances local prediction, while **multi-scale inference** (Section 4.4) captures objects at different scales. Scores from an existing global predictor can optionally be ensembled with LoCo's local scores.

4.1 Preliminaries

We first revisit the default CLIP pipeline for multi-label recognition. Given an image x, the ViT-based visual encoder E_{v} divides it into $H \times W$ patches and outputs a sequence of visual tokens

$$[\mathbf{v}_{\mathrm{cls}}, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{HW}] = E_{\mathrm{v}}(x), \tag{1}$$

where $\mathbf{v}_{\mathrm{cls}}$ is the global `[CLS]` token and $\mathbf{v}_i$ denotes the i-th local patch token. Given a category name set $\mathcal{S} = \{s_1, s_2, \ldots, s_C\}$ for classes $1, 2, \ldots, C$, each category name is converted into a textual prompt $\mathrm{Prompt}(\cdot)$ and encoded by the CLIP text encoder E_{t} as

$$\mathbf{q}_c = E_{\mathrm{t}}(\mathrm{Prompt}(s_c)), \quad c = 1, 2, \ldots, C. \tag{2}$$

The default CLIP pipeline uses only the global token to compute the image-level multi-label prediction score via cosine similarity

$$\hat{\mathbf{y}}^{\mathrm{g}} = [\langle \mathbf{v}_{\mathrm{cls}}, \mathbf{q}_1 \rangle, \langle \mathbf{v}_{\mathrm{cls}}, \mathbf{q}_2 \rangle, \ldots, \langle \mathbf{v}_{\mathrm{cls}}, \mathbf{q}_C \rangle]. \tag{3}$$

4.2 Spatially Consistent Visual Encoder

Following the analysis in Section 3.2, we adopt an existing spatially consistent encoder instead of training a new one. CLIP's last ViT block is a major source of local spatial inconsistency: its query–key attention mixes information across regions, and the final residual connection and FFN further propagate global context into patch tokens. Existing training-free corrections modify these operations to improve spatial consistency, *e.g.*, SCLIP [18] replaces the final-layer query–key attention with query–query and key–key correlations while keeping the residual and FFN paths, whereas ClearCLIP [31] uses self-self attention and drops the final residual connection and FFN.

In our implementation, we adopt SC-CLIP [19] as LoCo's spatially consistent encoder. Specifically, SC-CLIP applies intermediate-layer similarity for self-calibration and multi-level feature fusion. For efficiency, we disable only its anomaly-token resolving module, which refines segmentation boundaries but introduces extra inference cost. All encoder operations are inherited from SC-CLIP and introduce no training or learnable parameters, and LoCo applies Selective Softmax to the resulting dense logits. Formally, the spatially consistent visual backbone E'_{v} outputs

$$[\mathbf{v}_{\mathrm{cls}}, \mathbf{v}'_1, \mathbf{v}'_2, \ldots, \mathbf{v}'_{HW}] = E'_{\mathrm{v}}(x), \tag{4}$$

where $\mathbf{v}'_i$ denotes the i-th spatially consistent local patch token.

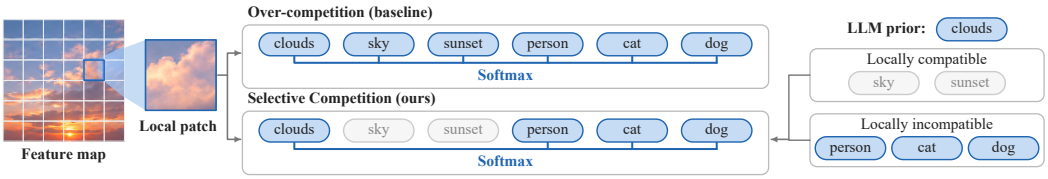


Figure 6: Selective local competition guided by an LLM compatibility prior, where *clouds* is selected as the target category.

4.3 Selective Local Competition

Given spatially consistent local features, each local region produces a category logit vector through vision-language matching. Let

$$\mathbf{s}_i = [\langle \mathbf{v}'_i, \mathbf{q}_1 \rangle, \langle \mathbf{v}'_i, \mathbf{q}_2 \rangle, \ldots, \langle \mathbf{v}'_i, \mathbf{q}_C \rangle] \tag{5}$$

denote the local logits on the i-th local token. A class-independent prediction treats all categories separately and therefore lacks relative competition, while a standard softmax over all categories assumes full mutual exclusiveness and may incorrectly suppress compatible concepts. To address this issue, we introduce selective local competition, where each category only competes with categories that are incompatible with it, as shown in Figure 6.

In open-vocabulary multi-label recognition, the category set may contain concepts of different granularities and abstraction levels. Ignoring such concept relations can either place compatible concepts into unnecessary competition or leave mutually exclusive concepts completely independent. For example, *sky* and *cloud* can coexist as a background region and its visible component. Rather than explicitly distinguishing all possible relation types, we simplify these relationships to **local compatibility**. Concepts that can jointly describe the same local region are treated as compatible, while the others are regarded as competitors.

To obtain such compatibility knowledge, we use an LLM to perform explicit semantic reasoning over the category name set, whose details are shown in Appendix D. Although CLIP is effective at measuring the similarity between an image region and a text query, it does not directly provide semantic compatibility or incompatibility relations among category names. In contrast, LLMs encode richer conceptual knowledge and are better suited for reasoning over category relations. Therefore, we feed the complete category name set $\mathcal{S} = \{s_1, s_2, \ldots, s_C\}$ into the LLM with a structured prompt, asking which categories can coexist with each category within the same local region.

Specifically, we parse the structured response offline and apply symmetrization: two categories are connected whenever either directed compatibility list includes the other. The resulting local compatibility graph is reused for all images with that category vocabulary. Its neighbors define the compatible set $\mathcal{A}_c^+ \subseteq \{1, \ldots, C\}$ for each category c. We then define the competition set of s_c as all categories except itself and its compatible categories

$$\mathcal{A}_c^- = \{j \mid j \neq c,\ j \notin \mathcal{A}_c^+\}. \tag{6}$$

Selective Softmax recalibrates the local logits by applying softmax to each category together with its incompatible competitors. For category c and the i-th token, its selective competition score is

$$\hat{y}^{\mathrm{l}}_{i,c} = \mathrm{SelSoftmax}_c\left(\mathbf{s}_i\right) = \frac{\exp\left(\mathbf{s}_{i,c}/\tau\right)}{\exp\left(\mathbf{s}_{i,c}/\tau\right) + \sum_{c' \in \mathcal{A}_c^-} \exp\left(\mathbf{s}_{i,c'}/\tau\right)}, \tag{7}$$

where τ is a temperature parameter. The score $\hat{y}^{\mathrm{l}}_{i,c}$ measures the relative advantage of category s_c relative to its incompatible competitors. Thus, Selective Softmax introduces competition where mutual exclusiveness is expected, while preserving compatibility among concepts.

4.4 Multi-scale Inference

LoCo performs multi-scale inference using grid cropping with 1×1, 2×2, and 3×3 partitions, yielding $K = 14$ image crops. Each crop is resized to 224×224 pixels and processed by the dense visual encoder to obtain a local logit map $\mathbf{s}_w \in \mathbb{R}^{H \times W \times C}$. Each map is then bilinearly interpolated to the corresponding crop region on a common $H \times W$ grid spanning the original image and positioned

according to the crop offset. The aligned logits are averaged at each grid location:

$$\bar{s}_{u,c} = \frac{1}{|\mathcal{W}(u)|} \sum_{w \in \mathcal{W}(u)} s^{\text{aligned}}_{w,u,c}, \tag{8}$$

where $\mathcal{W}(u)$ denotes the set of crops covering location u. This produces a merged logit map $\bar{\mathbf{s}} \in \mathbb{R}^{H \times W \times C}$. We then apply Selective Softmax to the merged logits, followed by spatial max pooling to obtain image-level local predictions:

$$\hat{y}^{\text{l}}_c = \max_u \text{SelSoftmax}_c\left(\bar{\mathbf{s}}_u\right), \tag{9}$$

where $\bar{\mathbf{s}}_u = [\bar{s}_{u,1}, \ldots, \bar{s}_{u,C}]$. While local scores capture region-level semantics, labels such as scenes, attributes, and contextual concepts still benefit from image-level representations. LoCo can optionally be combined with an existing global predictor g. Applying g to all K crops yields one global score vector per crop, $\hat{\mathbf{y}}^{\text{g}}_w = g(x_w, \mathcal{S})$, and the image-level global score is their unweighted mean:

$$\hat{\mathbf{y}}^{\text{g}} = \frac{1}{K} \sum_{w=1}^{K} \hat{\mathbf{y}}^{\text{g}}_w. \tag{10}$$

The final prediction is obtained by

$$\hat{\mathbf{y}} = \alpha \hat{\mathbf{y}}^{\text{g}} + (1 - \alpha)\hat{\mathbf{y}}^{\text{l}}, \tag{11}$$

where α weights the global prediction. LoCo thus serves as a local branch to which existing global OV-MLR predictors can be attached.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate our method on four widely used multi-label recognition datasets: Pascal VOC 2007 (VOC), MS-COCO 2014 (COCO), NUS-WIDE (NUS), and Visual Genome 150 (VG). Performance on the test sets is reported. VOC [36] and COCO [37] have segmentation-level annotations and thus exclusive categories, whereas NUS [38] and VG [39, 40] contain more complex non-exclusive categories, which are shown in Table 1. These benchmarks cover 20–150 categories and different image-level label densities.

Baselines. We compare LoCo with state-of-the-art OV-MLR methods: the original CLIP [1]; text-prompt- or text-prototype-based methods including TaI-DPT [8], TaI-Adapter [17], CoMC [10], and DualPrompt [14]; the local-prediction-based method TagCLIP [6]; and the composition-based method SPARC [26].

Evaluation metrics. We use mean average precision (mAP) and F1-score, macro-averaged across categories, as the main evaluation metrics. **mAP** measures the ranking quality of category predictions, which is widely adopted in previous literature [8, 12, 13, 6]. **F1-score** evaluates the binarized prediction after thresholding, which quantifies the practical discriminability of multi-label prediction. For each method and dataset, we select the threshold maximizing validation F1-score over 0.01–0.99 (step 0.01) and apply it to the test set. Component ablations also report **Recall@10**, the fraction of annotated positives retrieved among each image's 10 highest-scoring categories, and micro **PR-AUC**, computed over image–category pairs within each dataset, to fully evaluate their capability.

Implementation details. For fair comparison, all methods adopt ViT-B/16 as the visual backbone and are evaluated with the same category sets. For text-as-image methods, the same LLM-generated caption datasets are used for text-as-image training. Each caption dataset has 40,000 captions and 2∼5 labels per sample. LoCo uses a Selective Softmax temperature of $\tau = 1/40$, multi-scale inference using grid cropping with 1×1, 2×2, and 3×3 partitions, and $\alpha = 1/3$ for optional global ensembling. Unless otherwise specified, results are obtained under the same input resolution of 224×224 and the same text prompt templates. As an exception, TagCLIP uses an input resolution of 448×448 and introduces additional background categories following its original setting. All experiments are conducted on a server equipped with NVIDIA GeForce RTX 3090 $\times$ 8 GPUs and an AMD EPYC 7K62 48-Core Processor.

Table 2: Experimental results on open-vocabulary multi-label recognition. mAP measures test-set ranking quality. *Avg.* is the unweighted mean over four datasets. **Best** and second-best results are highlighted.

Method	Venue	F1-score (%)					mAP (%)				
		VOC	COCO	NUS	VG	*Avg.*	VOC	COCO	NUS	VG	*Avg.*
CLIP	ICML'21	65.2	39.0	30.8	21.1	39.0	79.1	55.0	40.6	31.2	51.5
TaI-DPT*	CVPR'23	75.4	55.2	41.8	37.6	52.5	86.1	66.0	46.2	41.2	59.9
TaI-Adapter*	arXiv'23	77.6	55.1	31.7	33.3	49.4	87.1	61.3	45.0	35.5	57.2
CoMC*	ICML'24	76.8	55.5	34.2	37.2	50.9	88.3	66.8	47.1	38.6	60.2
TagCLIP	AAAI'24	**85.3**	61.6	35.8	37.8	55.1	92.5	72.3	41.7	42.5	62.2
SPARC	CVPR'25	72.7	53.9	33.5	34.0	48.5	87.6	66.2	46.5	39.0	59.8
DualPrompt†	ICLR'26	76.0	57.8	38.4	37.2	52.4	87.8	69.5	48.1	36.5	60.5
LoCo	Ours	83.1	65.1	40.7	45.0	58.5	**93.0**	76.6	45.3	49.2	66.0
LoCo + TaI-DPT*	Ours	84.2	65.7	43.0	45.8	59.7	92.5	77.4	46.4	49.3	66.4
LoCo + DualPrompt†	Ours	84.1	**66.3**	**43.3**	**45.9**	**59.9**	**93.0**	**78.3**	**48.2**	**49.9**	**67.4**

* Extra training required.
† Extra annotations required.

Table 3: Ablation of the standalone local branch. *Sel.* denotes Selective Softmax and *M.S.* denotes 14-crop multi-scale inference. **Best** and second-best are highlighted.

Backbone	*Sel.*	*M.S.*	*Avg.* (%)			
			F1-score	mAP	Recall @10	PR-AUC
CLIP			46.4	55.0	71.6	43.1
ClearCLIP			52.5	59.8	79.8	51.0
ClearCLIP	✓		53.8	60.0	81.8	54.0
ClearCLIP		✓	54.7	63.7	81.6	53.7
ClearCLIP	✓	✓	56.0	64.0	83.5	56.8
SC-CLIP			54.9	62.4	82.2	54.7
SC-CLIP	✓		56.7	62.9	83.5	57.8
SC-CLIP		✓	56.6	65.5	83.9	56.6
SC-CLIP	✓	✓	**58.5**	**66.0**	**85.1**	**59.8**

Table 4: Fair comparison between LoCo and baseline method TaI-DPT for local prediction with the same dense encoder (Backbone) and the same multi-scale inference (*M.S.*).

Backbone	Method	*M.S.*	*Avg.* (%)			
			F1-score	mAP	Recall @10	PR-AUC
CLIP	TaI-DPT		52.5	59.4	75.2	47.1
CLIP	TaI-DPT	✓	51.1	60.2	75.6	47.9
SC-CLIP	TaI-DPT		55.8	63.1	81.1	54.6
SC-CLIP	TaI-DPT	✓	55.7	64.8	81.1	54.7
CLIP	LoCo		45.9	52.0	72.4	43.6
CLIP	LoCo	✓	47.8	55.3	74.4	47.0
SC-CLIP	LoCo		56.7	62.9	83.5	57.8
SC-CLIP	LoCo	✓	**58.5**	**66.0**	**85.1**	**59.8**

5.2 Comparison Results

Table 2 compares LoCo with state-of-the-art methods on four benchmark datasets. Standalone LoCo achieves the highest average F1-score and mAP among the standalone methods, with average gains of 3.4% in F1-score and 3.8% in mAP over TagCLIP. When ensembled with global predictions from TaI-DPT, LoCo achieves 59.7% F1-score and 66.4% mAP. LoCo + DualPrompt achieves the highest averages of 59.9% F1-score and 67.4% mAP, with gains of 4.8% and 5.2% over TagCLIP. The consistent gains across diverse datasets validate the proposed local competition mechanism and spatially consistent local representations in enhancing multi-label discriminability.

5.3 Ablation Study

Visual backbone matters. Table 3 evaluates the contributions of LoCo's components. Under Softmax and single-crop inference, replacing CLIP with ClearCLIP and SC-CLIP improves average F1-score from 46.4% to 52.5% and 54.9%, respectively, and mAP from 55.0% to 59.8% and 62.4%. These improvements highlight the importance of spatially consistent local features for local prediction.

Selective Softmax consistently contributes. Table 3 shows that, across ClearCLIP and SC-CLIP with single- and multi-scale inference, Selective Softmax improves average F1-score by 1.3–1.9% and mAP by 0.3–0.5%. It also improves Recall@10 by 1.2–2.0% and PR-AUC by 3.0–3.2%. These consistent gains across four evaluation metrics support the effectiveness of selective local competition in improving multi-label discriminability. Per-dataset results are provided in Appendix B.1.

Multi-scale inference further helps. Across different visual backbones (ClearCLIP or SC-CLIP) and local competition conditions (with and without Selective Softmax), multi-scale inference improves average F1-score by 1.7–2.2%, mAP by 3.1–4.0%, Recall@10 by 1.6–1.8%, and PR-AUC by 1.9–2.8%, demonstrating the benefit of capturing local evidence at different scales.

Table 5: Comparison of different compatibility priors. GPT-5.5 Prior is finally adopted.

Rule / Prior	F1-score	mAP
Raw logits	53.1	62.8
Softmax	56.6	65.5
Random Prior	52.9	61.5
WordNet Prior	57.0	65.5
GPT-5.5 Prior	58.5	**66.0**
LLMs Consensus Prior	**58.6**	**66.0**

Table 6: Computation cost and efficiency of standalone methods. Size is the image's width and height; crop counts are specified for LoCo.

Method (size; crops)	mAP	FLOPs	Throughput	Peak Mem
CLIP (224; –)	51.5	35.1 G/img	380.9 img/s	0.8 GB
TaI-DPT (224; –)	59.9	35.3 G/img	377.6 img/s	0.8 GB
TagCLIP (224; –)	61.0	70.6 G/img	25.1 img/s	1.2 GB
TagCLIP (448; –)	62.2	228.4 G/img	8.4 img/s	9.3 GB
LoCo (224; 1)	62.9	35.2 G/img	342.1 img/s	1.6 GB
LoCo (224; 14)	66.0	493.4 G/img	27.8 img/s	5.8 GB

LoCo better exploits spatial consistency and inference budget. Table 4 compares LoCo with TaI-DPT under the same visual backbone and inference settings for local prediction. TaI-DPT benefits from both components: replacing CLIP with SC-CLIP increases average mAP from 59.4% to 63.1%, and multi-scale inference further raises it to 64.8%. With SC-CLIP and multi-scale inference, LoCo outperforms TaI-DPT across all four metrics. These results show that LoCo makes more effective use of spatially consistent local features and multi-scale inference than TaI-DPT.

5.4 Local Compatibility Prior Analysis

Non-competition and over-competition harm. Table 5 examines the need for selective competition and the choice of compatibility prior under the same local model and 14-crop inference. Selective Softmax with the GPT-5.5 prior achieves 58.5% F1-score and 66.0% mAP, outperforming both raw logits without competition (53.1% and 62.8%) and full Softmax with excessive competition (56.6% and 65.5%). This comparison supports the necessity of selectively imposing competition among locally incompatible categories.

LLM generates reliable category compatibility. Table 5 compares four compatibility prior strategies. The random prior randomly assigns compatibility between category pairs, while the WordNet prior derives compatibility from is-a, part-of, and synonym relations. The GPT-5.5 prior uses a compatibility graph generated by GPT-5.5, whereas the LLM consensus takes the intersection of graphs generated independently by four different LLMs. GPT-5.5 outperforms both random and WordNet priors in F1-score and mAP, supporting LLM-based compatibility construction. The consensus yields a marginally higher F1-score (58.6% vs. 58.5%) with the same mAP (66.0%). We adopt the single GPT-5.5 prior for simplicity, given its comparable performance without combining multiple LLMs. Details are provided in Appendices D.2 and D.4.

5.5 Computational Cost Analysis

Table 6 shows that LoCo without *M.S.* matches the original CLIP in FLOPs and throughput, whereas *M.S.* inference raises its cost by about 14 times due to grid cropping, giving an **accuracy-efficiency trade-off**. Despite higher total FLOPs, LoCo with *M.S.* still keeps a throughput advantage over TagCLIP at high resolution, because its crop-level computation is fully parallelizable while TagCLIP relies on a less hardware-friendly two-stage pipeline. The trade-off is acceptable, since high-resolution recognition is often necessary for small objects.

6 Conclusion

This paper revisits the role of local prediction in OV-MLR. We show that local prediction becomes a strong discriminative branch when local visual features are spatially consistent and category competition is aware of local semantic compatibility. Motivated by the observation of local semantic sparsity, we propose LoCo, a selective local competition framework that uses spatially consistent dense features and Selective Softmax to impose competition only among locally incompatible categories. Experiments on four multi-label benchmarks show that standalone LoCo improves average F1-score and mAP over TagCLIP by 3.4% and 3.8%, respectively. With the optional DualPrompt global predictor, LoCo achieves the highest averages among the compared methods, improving over TagCLIP by 4.8% in F1-score and 5.2% in mAP. Beyond this framework, our findings suggest that local semantic sparsity offers a useful basis for OV-MLR and motivate future work on local semantic structure and dense vision-language representations.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China under Grants U24A20322 and 62576094. This research work was also supported by the Big Data Computing Center of Southeast University. The authors declare no competing interests.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[2] Sarah M. Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15645–15655, 2023.

[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

[4] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 30662–30680, 2023.

[5] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W. Tsang. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(11): 7955–7974, 2022.

[6] Yuqi Lin, Minghao Chen, Kaipeng Zhang, Hengjia Li, Mingming Li, Zheng Yang, Dongqin Lv, Binbin Lin, Haifeng Liu, and Deng Cai. Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of CLIP without training. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3513–3521, 2024.

[7] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[8] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2808–2817, 2023.

[9] Xiangyu Wu, Qing-Yuan Jiang, Yang Yang, Yi-Feng Wu, Qing-Guo Chen, and Jianfeng Lu. TAI++: text as image for multi-label image classification by co-learning transferable prompt. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5226–5234, 2024.

[10] Yicheng Liu, Jie Wen, Chengliang Liu, Xiaozhao Fang, Zuoyong Li, Yong Xu, and Zheng Zhang. Language-driven cross-modal classifier for zero-shot multi-label image recognition. In *Forty-first International Conference on Machine Learning (ICML)*, pages 32173–32183, 2024.

[11] Shuo Yang, Zirui Shang, Yongqi Wang, Derong Deng, Hongwei Chen, Xinxiao Wu, and Qiyuan Cheng. Image-free multi-label image recognition via llm-powered hierarchical prompt tuning. *Pattern Recognition (PR)*, 174:112986, 2026.

[12] Chun-Mei Feng, Kai Yu, Xinxing Xu, Salman Khan, Rick Siow Mong Goh, Wangmeng Zuo, and Yong Liu. Text to image for multi-label image recognition with joint prompt-adapter learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 47(9): 7660–7674, 2025.

[13] Dong Un Kang, Hyunwoo Lee, and Se Young Chun. Class concept representation from contextual texts for training-free multi-label recognition. *IEEE Access*, 13:36965–36977, 2025.

[14] Ming-Kun Xie, Zhiqiang Kou, Zhongnian Li, Gang Niu, and Masashi Sugiyama. Unlocking the power of co-occurrence in clip: A dualprompt-driven method for training-free zero-shot multi-label classification. In *International Conference on Learning Representations (ICLR)*, 2026.

[15] Sanath Narayan, Akshita Gupta, Salman H. Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah. Discriminative region-based multi-label zero-shot learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8711–8720, 2021.

[16] Samyak Rawlekar, Yujun Cai, Yiwei Wang, Ming-Hsuan Yang, and Narendra Ahuja. Disentangling CLIP features for enhanced localized understanding. *arXiv preprint arXiv:2502.02977*, 2025.

[17] Xuelin Zhu, Jiuxin Cao, Jian Liu, Dongqi Tang, Furong Xu, Weijia Liu, Jiawei Ge, Bo Liu, Qingpei Guo, and Tianyi Zhang. Text as image: Learning transferable adapter for multi-label classification. *arXiv preprint arXiv:2312.04160*, 2023.

[18] Feng Wang, Jieru Mei, and Alan L. Yuille. SCLIP: rethinking self-attention for dense vision-language inference. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *European Conference on Computer Vision (ECCV)*, pages 315–332, 2024.

[19] Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, Yansong Tang, Jie Zhou, and Jiwen Lu. Self-calibrated CLIP for training-free open-vocabulary segmentation. *IEEE Transactions on Image Processing (IEEE TIP)*, 34:8271–8284, 2025.

[20] Bo Han, Zhuoming Li, Xiaoyu Wang, Yaxin Hou, Hui Liu, Junhui Hou, and Yuheng Jia. Dicap: Distribution-calibrated pseudo-labeling for semi-supervised multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 21540–21548, 2026.

[21] Zhuoming Li, Yuheng Jia, Mi Yu, and Zicong Miao. Calibrated disambiguation for partial multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 18620–18628, 2025.

[22] Fuchao Yang, Yuheng Jia, Hui Liu, Yongqiang Dong, and Junhui Hou. Noisy label removal for partial multi-label learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 3724–3735, 2024.

[23] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. CDUL: clip-driven unsupervised learning for multi-label image classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1348–1357, 2023.

[24] Shouwen Wang, Qian Wan, Junbin Gao, and Zhigang Zeng. Language-driven multi-label zero-shot learning with semantic granularity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1968–1978, 2025.

[25] Shuo Yang, Zirui Shang, Yongqi Wang, Derong Deng, Hongwei Chen, Qiyuan Cheng, and Xinxiao Wu. Data-free multi-label image recognition via llm-powered prompt tuning. *arXiv preprint arXiv:2403.01209*, 2024.

[26] Kevin Miller, Aditya Gangrade, Samarth Mishra, Kate Saenko, and Venkatesh Saligrama. SPARC: score prompting and adaptive fusion for zero-shot multi-label recognition in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4313–4321, 2025.

[27] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15305–15314, 2023.

[28] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision (ECCV)*, pages 696–712, 2022.

[29] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025.

[30] Osman Ülger, Maksymilian Kulicki, Yuki Asano, and Martin R. Oswald. Auto-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24266–24275, 2025.

[31] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing CLIP representations for dense vision-language inference. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *European Conference on Computer Vision (ECCV)*, pages 143–160, 2024.

[32] Monika Wysoczanska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzcinski, and Patrick Pérez. Clip-dinoiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *European Conference on Computer Vision (ECCV)*, pages 320–337, 2024.

[33] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves CLIP for open-vocabulary segmentation. In *European Conference on Computer Vision (ECCV)*, pages 70–88, 2024.

[34] Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to DINO: bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22025–22035, 2025.

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal (TMLR)*, 2024.

[36] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.

[37] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[38] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In Stéphane Marchand-Maillet and Yiannis Kompatsiaris, editors, *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, 2009.

[39] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

[40] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3106, 2017.

Table of Contents for the Appendix

A Text Correlation and Response Separability **15**

B Supplementary Ablation Results **15**

B.1 Per-Dataset Ablation Results 15

B.2 Temperature Sensitivity 15

C Matched Encoder and Inference Comparisons **16**

D Compatibility Prior Analysis **17**

D.1 Prior Construction 17

D.2 Generation, Prompt, and Cross-LLM Stability 17

D.3 Recognition Sensitivity 18

D.4 WordNet-Derived Compatibility Relations 18

E Qualitative Results and Failure Cases **18**

E.1 More qualitative results 18

E.2 Failure cases 19

F Limitations and Future Work **20**

A Text Correlation and Response Separability

Table 7 reports mean pairwise cosine similarity among category text prototypes for the four benchmark vocabularies. CLIP has higher mean pairwise similarity than TaI-DPT in each vocabulary. These statistics describe text-prototype geometry and complement the response-distribution analysis in Table 8.

Table 7: Mean text-prototype cosine similarities, displayed in units of 10^{-2}. Multiply displayed similarities by 10^{-2} to recover the dimensionless cosine values.

Dataset	Classes	CLIP mean	TaI-DPT mean
VOC	20	81.0	24.0
COCO	80	76.0	14.0
NUS	81	80.0	19.0
VG	150	78.0	17.0

Table 8 summarizes the reported response distributions across the four datasets, separating positive-label and negative-label scores. Raw scores use different scales across methods. Standardized separation is reported using Cohen's D:

$$D = \frac{\mu^+ - \mu^-}{s_{\text{pool}}}, \tag{12}$$

where μ^+ and μ^- are the positive and negative score means, and s_{pool} is the pooled estimated population standard deviation. The reported D values are 1.5 for CLIP, 2.3 for TaI-DPT, and 4.1 for LoCo. These are descriptive comparisons of complete model predictions.

Table 8: Reported response statistics across four datasets. All score statistics are displayed in units of 10^{-3}; multiply them by 10^{-3} to recover the original response scale. Cohen's D is dimensionless and unscaled.

Model	Positive only Mean $\pm$ Std.	Negative only Mean $\pm$ Std.	Cohen's D
CLIP	175.0 ± 32.0	138.0 ± 24.0	1.5
TaI-DPT	17.0 ± 5.0	11.0 ± 3.0	2.3
LoCo	449.0 ± 248.0	55.0 ± 78.0	4.1

B Supplementary Ablation Results

B.1 Per-Dataset Ablation Results

Table 9 compares Softmax and Selective Softmax across ClearCLIP and SC-CLIP with one or 14 crops. The table reports F1-score, mAP, micro PR-AUC, and Recall@10 (all in %). Recall@10 is evaluated against annotated positive labels. Averages are unweighted means over the four datasets; **best** and second-best ranks use unrounded scores.

Selective Softmax improves the reported dataset-averaged metrics in all four backbone/crop configurations. The individual results include a slight decline in VG micro PR-AUC for 14-crop ClearCLIP: both values round to 40.5%. This metric-specific exception is distinct from the VG F1-score improvements in Table 9.

B.2 Temperature Sensitivity

We analyze the sensitivity of the Selective Softmax temperature τ in Figure 7. Overall, the best results are achieved at $\tau = 1/40$ on 3/4 benchmarks, while on the remaining dataset $\tau = 1/20$ is only slightly better than $\tau = 1/40$. This suggests that the proposed method is not highly sensitive to the exact temperature choice around this range. In particular, the mAP performance remains relatively stable near $\tau = 1/40$, indicating that LoCo is robust to moderate changes of the temperature parameter. Therefore, we use $\tau = 1/40$ as the default setting in all experiments.

Table 9: Per-dataset ablation of the standalone local branch. Top: F1-score and mAP. Bottom: micro PR-AUC and Recall@10. *Sel.* denotes Selective Softmax and *M.S.* denotes 14-crop inference; unchecked settings use Softmax and one crop, respectively.

Backbone	*Sel.*	*M.S.*	F1-score (%)					mAP (%)				
			VOC	COCO	NUS	VG	*Avg.*	VOC	COCO	NUS	VG	*Avg.*
ClearCLIP			76.2	60.7	34.8	38.2	52.5	88.3	70.8	37.9	42.0	59.8
ClearCLIP	✓		78.1	61.0	36.2	39.9	53.8	88.5	70.6	38.4	42.7	60.0
ClearCLIP		✓	77.9	65.0	35.2	40.8	54.7	91.2	**77.1**	40.2	46.4	63.7
ClearCLIP	✓	✓	80.0	**65.1**	36.6	42.4	56.0	91.3	76.9	40.8	46.9	64.0
SC-CLIP			79.6	60.7	39.2	40.0	54.9	90.8	71.1	43.1	44.4	62.4
SC-CLIP	✓		82.0	61.3	**41.0**	42.5	56.7	91.1	70.9	44.2	45.5	62.9
SC-CLIP		✓	79.9	64.7	39.0	42.6	56.6	92.8	76.8	44.3	48.2	65.5
SC-CLIP	✓	✓	**83.1**	65.1	40.7	**45.0**	**58.5**	**93.0**	76.6	**45.3**	**49.2**	**66.0**

Backbone	*Sel.*	*M.S.*	PR-AUC (%)					Recall@10 (%)				
			VOC	COCO	NUS	VG	*Avg.*	VOC	COCO	NUS	VG	*Avg.*
ClearCLIP			75.0	60.0	31.9	37.0	51.0	99.6	85.7	82.6	51.3	79.8
ClearCLIP	✓		79.7	63.0	36.1	37.2	54.0	99.7	89.5	85.6	52.3	81.8
ClearCLIP		✓	76.8	65.0	32.5	40.5	53.7	99.9	88.6	83.4	54.5	81.6
ClearCLIP	✓	✓	81.6	67.8	37.3	40.5	56.8	99.9	**92.3**	86.3	55.4	83.5
SC-CLIP			80.1	62.6	37.0	39.1	54.7	99.6	88.4	87.8	53.2	82.2
SC-CLIP	✓		84.1	65.8	**41.5**	39.9	57.8	99.6	89.6	89.9	55.0	83.5
SC-CLIP		✓	80.8	66.6	36.8	42.3	56.6	99.9	90.8	88.7	56.2	83.9
SC-CLIP	✓	✓	**84.9**	**69.7**	41.4	**43.0**	**59.8**	**99.9**	92.1	**90.6**	**57.9**	**85.1**

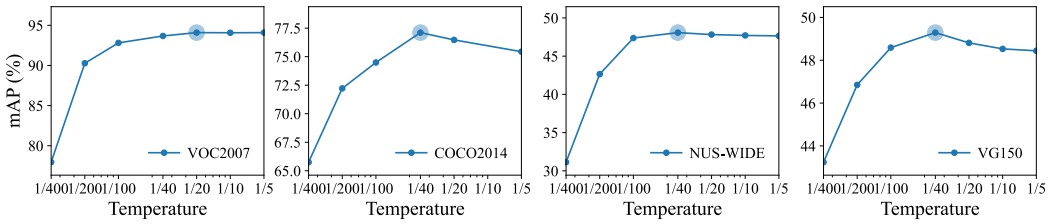


Figure 7: Sensitivity analysis of the hyperparameter τ.

C Matched Encoder and Inference Comparisons

Table 10 expands the matched comparison summarized in Table 4 to per-dataset values. It compares LoCo with TaI-DPT using the same SC-CLIP dense encoder, with and without *M.S.*, alongside the original CLIP results. All averages are unweighted means over the four datasets; scores are rounded to one decimal after calculation and highlighted ranks use unrounded values.

Table 10: Per-dataset matched comparison against TaI-DPT on the same SC-CLIP backbone, with and without *M.S.* Top: F1-score and mAP. Bottom: Recall@10 and micro PR-AUC. Recall@10 is evaluated against annotated positive labels.

Backbone	Method	*M.S.*	F1-score (%)					mAP (%)				
			VOC	COCO	NUS	VG	*Avg.*	VOC	COCO	NUS	VG	*Avg.*
CLIP	TaI-DPT		75.4	55.2	41.8	37.6	52.5	86.1	65.8	45.9	39.8	59.4
SC-CLIP	TaI-DPT		78.9	60.8	41.7	41.7	55.8	89.8	71.6	46.5	44.5	63.1
CLIP	TaI-DPT	✓	75.8	52.0	40.1	36.6	51.1	86.5	66.1	47.8	40.2	60.2
SC-CLIP	TaI-DPT	✓	78.7	58.6	**43.8**	41.8	55.7	91.4	71.9	**50.6**	45.3	64.8
SC-CLIP	LoCo		82.0	61.3	41.0	42.5	56.7	91.1	70.9	44.2	45.5	62.9
SC-CLIP	LoCo	✓	**83.1**	**65.1**	40.7	**45.0**	**58.5**	**93.0**	**76.6**	45.3	**49.2**	**66.0**

Backbone	Method	*M.S.*	Recall@10 (%)					PR-AUC (%)				
			VOC	COCO	NUS	VG	*Avg.*	VOC	COCO	NUS	VG	*Avg.*
CLIP	TaI-DPT		95.8	78.0	78.3	48.9	75.2	74.3	49.3	33.0	31.6	47.1
SC-CLIP	TaI-DPT		97.8	84.2	87.8	54.8	81.1	78.6	57.9	43.1	38.8	54.6
CLIP	TaI-DPT	✓	96.2	78.4	78.6	49.0	75.6	75.1	49.8	34.7	31.9	47.9
SC-CLIP	TaI-DPT	✓	97.7	84.4	87.6	54.6	81.1	78.8	57.5	**43.2**	39.3	54.7
SC-CLIP	LoCo		99.6	89.6	89.9	55.0	83.5	84.1	65.8	41.5	39.9	57.8
SC-CLIP	LoCo	✓	**99.9**	**92.1**	**90.6**	**57.9**	**85.1**	**84.9**	**69.7**	41.4	**43.0**	**59.8**

With the same SC-CLIP encoder and *M.S.*, LoCo achieves higher average F1-score, mAP, Recall@10, and PR-AUC than TaI-DPT. The gains indicate that LoCo better uses the spatially consistent local features under the matched inference setting.

D Compatibility Prior Analysis

D.1 Prior Construction

To obtain prior knowledge about local coexistence and mutual exclusion among categories in a given label set, we query an LLM with a manually designed prompt template. For each dataset, we provide the complete category name list and ask the LLM to determine, under a semantic segmentation setting, which category concepts can simultaneously apply to the same pixel and which categories should be regarded as mutually exclusive. In our experiments, we use the GPT-5.5 Thinking model as the LLM. The prompt requests the output in the form of a Python dictionary, `dict[str, list[str]]`, where each key denotes a category name and the corresponding value lists the categories that can coexist with it at the pixel level. The LLM returns a structured dictionary-like response, in which each category is associated with a list of semantically compatible categories. We parse the responses offline into a local compatibility graph and apply OR symmetrization: a pair is compatible if either category lists the other. The resulting graph is fixed and reused for images with the same category vocabulary. Compatibility describes concepts that can apply to the same local region, rather than merely co-occurring somewhere in one image. Each category competes with all categories except itself and its compatible neighbors. The prompt we use is given as follows.

```
In semantic segmentation scenarios, it is usually assumed that a pixel
can belong to only one concept from the category set, thus there are no
cases where multiple concepts can coexist. Now, given a category set
containing n category names, I would like to obtain a prior assumption
about which categories are mutually exclusive at the same pixel and
which category concepts can simultaneously apply to the same pixel.
Please first analyze the category set, and then generate a Python
dictionary in the form of dict[str, list[str]], listing the categories
that can coexist at the same pixel.
{{category_name_list}}
```

D.2 Generation, Prompt, and Cross-LLM Stability

We evaluate graph stability under three sources of variation: five independent GPT-5.5 generations, three semantically equivalent prompt variants, and four different LLMs. Each repeated generation uses a separate single-turn conversation; prompt comparisons vary the wording, and model comparisons use the same prompt. Table 11 separates repeated generation from prompt variation. "Shared" counts the edges present in every graph in that condition. For edge sets A and B, Jaccard similarity is $|A \cap B|/|A \cup B|$ and overlap is $|A \cap B|/\min(|A|, |B|)$; the tables report average pairwise similarities in percent.

Table 11: Compatibility-graph stability under five independent generations and three prompt variants. Edge ranges and shared-edge counts are integers; Jaccard and overlap are percentages.

Dataset	Five independent generations				Three prompt variants			
	Edges	Shared	Jaccard	Overlap	Edges	Shared	Jaccard	Overlap
VOC (20)	25–42	25	77.0%	97.6%	31–45	26	63.4%	88.4%
COCO (80)	213–283	174	73.1%	90.6%	190–221	151	71.8%	88.0%
NUS (81)	277–428	221	69.1%	90.0%	266–367	210	63.9%	87.3%
VG (150)	632–1019	531	66.7%	88.9%	593–822	390	52.3%	77.4%

Table 12 reports the cross-LLM edge-set comparisons. The four-model LLM consensus is the intersection of their generated compatibility graphs; its shared-edge counts are 22, 70, 125, and 286 for VOC, COCO, NUS, and VG. Cross-model Jaccard similarity ranges from 35.8% to 70.3%, whereas overlap ranges from 79.7% to 93.9%. These summarize graph structure, separately from recognition performance.

Table 12: Cross-LLM compatibility-graph consistency. Shared edges form the four-model LLM consensus. Similarities are reported in percent.

Dataset	Edges	Shared	Jaccard (%)	Overlap (%)
VOC (20)	25–38	22	70.3%	93.9%
COCO (80)	98–456	70	35.8%	88.6%
NUS (81)	157–597	125	40.9%	84.5%
VG (150)	438–882	286	48.9%	79.7%

D.3 Recognition Sensitivity

Table 13 compares recognition performance under the LLM-generated priors and their consensus. Average F1-score ranges from 58.1% to 58.6%, and mAP from 65.6% to 66.1%, despite differences in graph structure. The consensus has 125.8 edges per dataset on average, computed from the four shared-edge counts, and retains 58.6% F1-score and 66.0% mAP. The final row preserves the reported cross-prior standard deviation; it is not a confidence interval over repeated recognition runs.

Table 13: Sensitivity to compatibility priors. Edges and scores are averaged over the four datasets. Scores are percentages; **best** and second-best values include ties at the reported precision.

Prior	*Avg.* edges	*Avg.* F1-score	*Avg.* mAP
glm-5.2	490.0	58.1	66.0
gemini-3.6-flash	182.3	58.3	**66.1**
claude-opus-4.6	290.5	58.5	65.6
gpt-5.5	314.8	58.5	66.0
LLM consensus	125.8	**58.6**	66.0
Reported std.	–	$\pm$0.2	$\pm$0.2

D.4 WordNet-Derived Compatibility Relations

Table 14 compares the WordNet-derived graph with the LLM consensus under the current vocabularies and graph-construction rules for is-a, part-of, and synonym relations. The LLM consensus is the intersection across four LLM-generated graphs. The LLM–WordNet intersection is the further intersection of that consensus with the WordNet-derived graph; these are different sets. For the empty WordNet graph on VOC, overlap is displayed as 0.0% by the table's reporting convention.

Table 14: WordNet-derived relations and the four-model LLM consensus. Jaccard and overlap compare their edge sets.

Dataset	WordNet edges	LLM-consensus edges	Jaccard (%)	Overlap (%)
VOC (20)	0	22	0.0%	0.0%
COCO (80)	3	70	1.4%	33.3%
NUS (81)	34	125	20.5%	79.4%
VG (150)	323	286	39.7%	60.5%

On COCO, the two graphs contain 3 and 70 edges, with one edge in the LLM–WordNet intersection; 69 LLM-consensus edges are absent from the WordNet-derived graph. On NUS, the respective counts are 34 and 125, with 27 shared edges, leaving 98 LLM-consensus edges absent from the WordNet-derived graph and 7 WordNet edges outside the consensus. Reported differences include scene or functional pairs such as person–backpack and apple–bowl. These coverage observations concern the chosen construction rules and category vocabularies. The recognition comparison in Table 5 shows that the WordNet prior improves F1-score over full competition, while the LLM priors achieve higher average F1-score and mAP.

E Qualitative Results and Failure Cases

E.1 More qualitative results

Beyond the visualizations in the main paper, we consistently observe that LoCo produces more spatially concentrated and semantically cleaner local responses across different datasets, which is shown in Figure 8 and Figure 9. In particular, Selective Softmax allows compatible categories

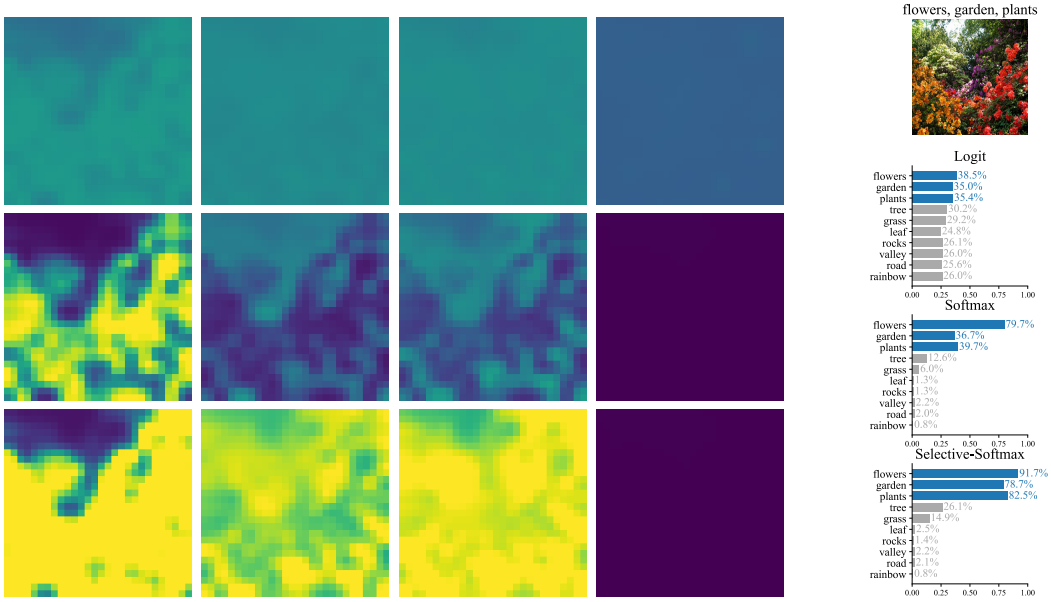


Figure 8: More qualitative results of LoCo's dense predictions. Class activation maps in rows 1, 2, and 3 show logits, Softmax predictions, and Selective Softmax predictions, while columns 1, 2, 3, and 4 indicate *flowers*, *garden*, *plants*, and *person* classes.

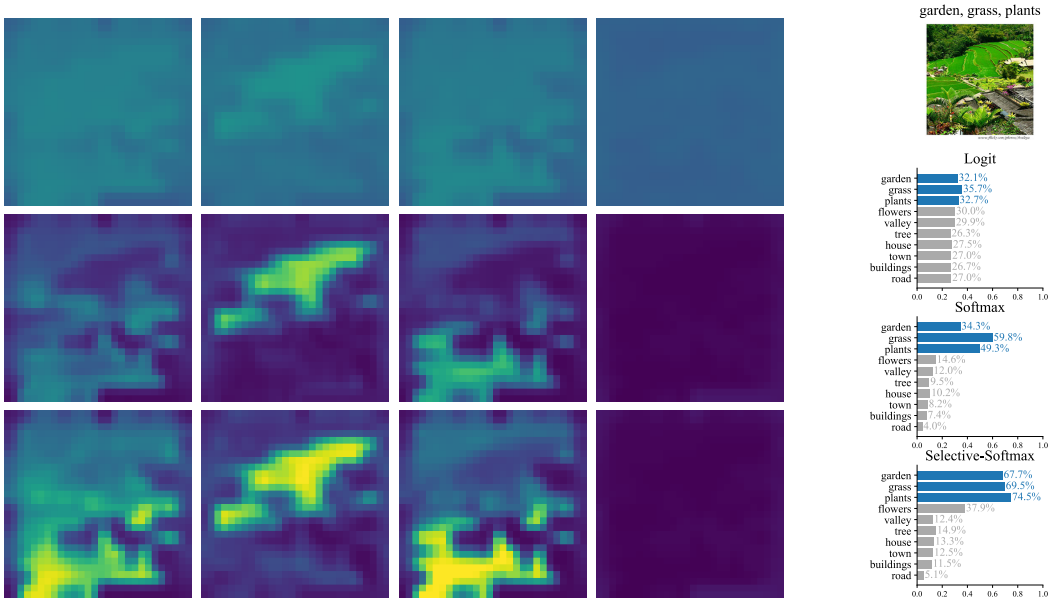


Figure 9: More qualitative results of LoCo's dense predictions. Class activation maps in rows 1, 2, and 3 show raw logits, softmax predictions, and selective softmax predictions, while columns 1, 2, 3, and 4 indicate *garden*, *grass*, *plants*, and *person* classes.

to remain simultaneously active on the same region, while still suppressing clearly incompatible concepts. This behavior is especially beneficial for category sets containing coarse-to-fine relations, part–whole composition, and other non-exclusive semantics, where standard softmax tends to over-suppress true positive labels.

E.2 Failure cases

Figure 10 illustrates three qualitative challenges: small or visually weak objects, categories requiring broader scene or attribute context, and ambiguous category boundaries. Some visually present

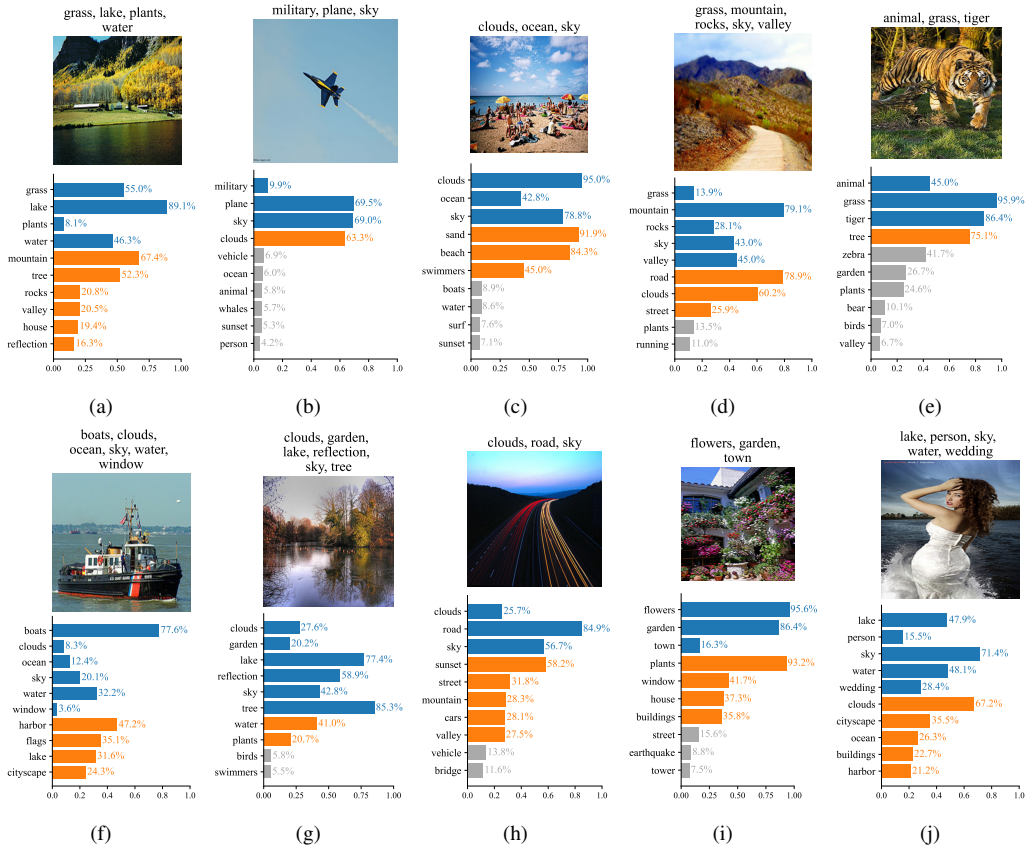


Figure 10: Failure cases for LoCo on the NUS dataset. Blue bars denote scores for annotated positive categories; orange bars highlight selected categories absent from the original annotations. The original annotations are shown above each image. Some unannotated concepts are visually present.

concepts are absent from the original annotations. These observations concern selected examples and inspected pairs; the benchmark comparisons retain the original annotations.

F Limitations and Future Work

LoCo uses a fixed binary compatibility graph for each category vocabulary. This prior approximates richer, context-dependent relations and requires compatibility relations for the target vocabulary. Context-adaptive or instance-conditioned compatibility is a direction for future work. Local multi-scale inference improves average recognition performance at additional computational cost; optional global ensembling adds a separate predictor's cost (Section 5.5).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 distinguish standalone LoCo from optional global fusion. Section 5 and appendix C report benchmark and matched comparisons; Appendix F describes the scope and limitations of the method.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of this work in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper presents an empirical study and includes no formal theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: Section 5 and appendix D.1 describe the method, datasets, hyperparameters, and LLM prompting procedure. The compatibility-graph versions used for the reported experiments and a complete reproduction protocol are not provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The supplementary implementation does not identify the compatibility-graph versions for the revised experiments or provide a complete reproduction protocol.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 and appendix D.1 specify the evaluation datasets, input settings, threshold selection, temperature, crop partitions, fusion weight, and compatibility-prior construction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars, confidence intervals, or significance tests for repeated recognition evaluations. Appendix D.2 reports graph-generation stability and cross-prior performance variation; these describe variation across compatibility graphs rather than confidence intervals for repeated recognition runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Section 5 identifies the server hardware, and Section 5.5 reports standalone inference measurements. Per-experiment execution times, total project compute, and the full measurement protocol are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms, in every respect, with the NeurIPS Code of Ethics. The authors have reviewed the NeurIPS Code of Ethics and have ensured that their research adheres to its guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper evaluates open-vocabulary recognition methods and does not include a societal-impact analysis.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work evaluates recognition methods on existing datasets and does not introduce a new image dataset or generative model.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: Original works are cited, but the manuscript does not explicitly list licenses and terms for all reused datasets, models, and code.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The paper describes the method and compatibility-prior construction, but complete release documentation for code and generated priors is not provided.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve research with human subjects requiring IRB or equivalent review.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Section 4 and appendix D.1 describe offline LLM construction of the fixed compatibility graph and provide the model name and prompt. Appendices D.2 and F discuss prior variation and the fixed binary approximation.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.