

Category-Adaptive Cross-Modal Semantic Refinement and Transfer for Open-Vocabulary Multi-Label Recognition

Haijing Liu, Tao Pu, Hefeng Wu, Keze Wang, Feng Gao, Fan Yang, Guanbin Li

Abstract—Benefiting from the generalization capability of CLIP, recent vision language pre-training (VLP) models have demonstrated the ability to capture a wide range of visual concepts in daily images. However, due to the presence of unseen categories in open-vocabulary settings, existing algorithms struggle to capture semantic correlations between categories, leading to suboptimal performance on open-vocabulary multi-label recognition (OV-MLR). Furthermore, the substantial variation in the number of discriminative areas across diverse object categories is misaligned with the fixed-number patch matching used in current methods, introducing noisy visual cues that hinder the capture of target semantics. To address these challenges, we propose a novel category-adaptive cross-modal semantic refinement and transfer (C²SRT) framework to model semantic correlations both within each category and across different categories, in a category-adaptive manner. The proposed framework consists of two complementary modules, i.e., intra-category semantic refinement (ISR) module and inter-category semantic transfer (IST) module. Specifically, the ISR module leverages the cross-modal knowledge of the VLP model to adaptively select a set of local discriminative regions that represent the semantics of the target category. The IST module adaptively discovers a set of correlated categories for a target category by constructing a category-adaptive correlation graph and transfers semantic knowledge from the correlated seen categories to unseen ones. Experiments on OV-MLR benchmarks demonstrate that the proposed C²SRT framework improves over current methods.

Index Terms—Multi-label image recognition, Open-vocabulary, Category-adaptive, Cross-modal, Vision-language model

I. INTRODUCTION

SINCE daily images inherently contain multiple semantic labels, multi-label recognition (MLR) [1], [2], [3], [4], [5], [6], which aims to identify target semantic labels in an input image, has garnered significant attention in the community. However, constrained by their predefined label space, these approaches often suffer significant performance degradation when classifying visual content from unseen categories (also referred as novel categories). To deal with this issue, recent works tend to study the task of open-vocabulary multi-label

This work was supported by National Natural Science Foundation of China (NSFC) under Grant No. 62272494 and 62322608, and Guangdong Basic and Applied Basic Research Foundation under Grant No. 2023A1515012845. (Corresponding author: Hefeng Wu)

Haijing Liu, Tao Pu, Hefeng Wu, and Keze Wang are with Sun Yat-sen University, Guangzhou 510006, China (e-mail: liuhj66@mail2.sysu.edu.cn, putao537@gmail.com, wuhefeng@gmail.com, kezewang@gmail.com).

Feng Gao and Fan Yang are with Peking University, Beijing 100871, China (e-mail: gaof@pku.edu.cn, fyang.eecs@pku.edu.cn).

Guanbin Li is with Sun Yat-sen University, Guangzhou 510006, China, and also with Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, China (e-mail: liguanbin@mail.sysu.edu.cn).

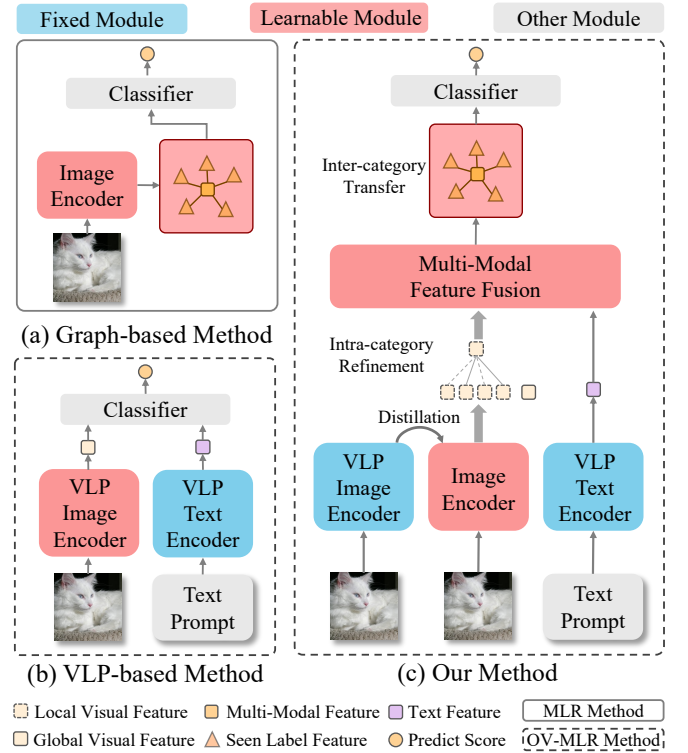


Fig. 1. Architectural differences between (a) traditional multi-label recognition methods and (b) open-vocabulary multi-label recognition methods. Compared with previous approaches, (c) our proposed method explores rich semantic correlation both within each category and across different categories.

recognition (OV-MLR) [7], [8], in which some target labels are unseen during the training phase. Compared with the traditional MLR, OV-MLR is more practical to real-world scenarios (e.g., autonomous driving [9], [10], scene understanding [11], [12], multi-modal recommendation [13], [14], and social media content annotation [15], [16]) because it requires models to generalize to novel categories that have not encountered before.

Due to the diverse appearance of objects within the same category, identifying the target semantic category based solely on visual input is challenging. Fortunately, as illustrated in Figure 2(a), strong semantic correlations across different categories can facilitate knowledge transfer from seen to unseen labels, enhancing the performance of semantic grounding. In traditional MLR, many prior works have introduced graph neural networks (GNNs) [17], [18], [19] to model inter-category

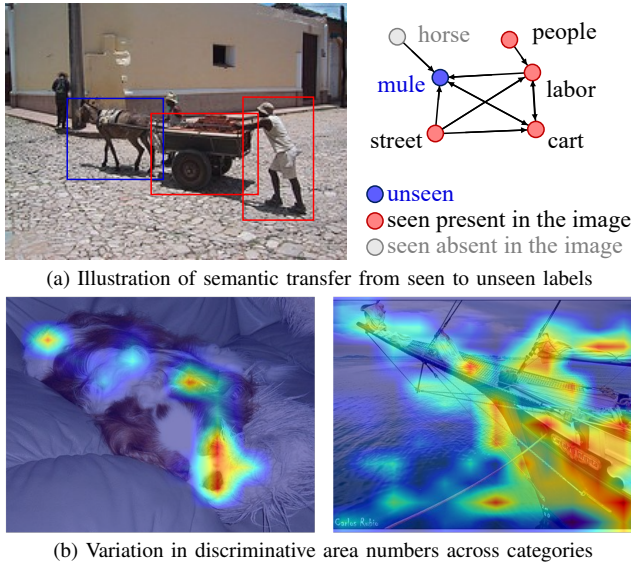


Fig. 2. Several examples of semantic correlations (a) across different categories and (b) within each category.

relationships, as shown in Figure 1(a). These approaches leverage prior knowledge such as statistical co-occurrence probabilities [20], [2], [21] and semantic similarities [6], [3] among categories to improve recognition accuracy. However, in open-vocabulary settings, novel labels hinder the accurate capture of co-occurrence information, posing a challenge for traditional MLR models in adapting to OV-MLR tasks. Consequently, while semantic similarities derived from textual embeddings may not accurately reflect complex semantic correlations, current OV-MLR models [7] rely primarily on textual embeddings to identify target categories. On the other hand, current OV-MLR methods leverage vision-language pre-training (VLP) models, as shown in Figure 1(b), to focus on local features. This approach, which has been widely validated as a key component in classical MLR, involves selecting a fixed number of patch features extracted by the VLP’s Image Encoder (such as ViT [22], [23]), thereby introducing discriminative regions into the visual features. However, this method ignores the substantial variation in the number of discriminative areas across different semantic categories, as presented in Figure 2(b). As a result, these algorithms achieve only suboptimal performance.

In this work, we propose a novel category-adaptive cross-modal semantic refinement and transfer (C²SRT) framework to effectively explore semantic correlations within and between categories in open-vocabulary scenarios. This framework consists of two complementary modules that adaptively refine intra-category discriminative regions and transfer inter-category semantic correlations. The C²SRT framework is built upon a VLP model with a learnable vision encoder that distills knowledge from the fixed vision encoder of the VLP. An intra-category semantic refinement (ISR) module is introduced to adaptively select semantically relevant local regions, thereby reducing the noise caused by object size and appearance variations. The ISR module quantifies the alignment between local features and the textual features of each category, adap-

tively selecting discriminative regional features as relevant visual representations. Furthermore, an inter-category semantic transfer (IST) module is designed to capture complex semantic correlations between categories, including unseen labels, thereby enhancing generalization capabilities in open-vocabulary scenarios. By leveraging external semantic priors (e.g., similarity-based or LLM-based relation mining), the IST module adaptively constructs a category correlation graph, enabling the transfer of semantic knowledge from correlated seen categories to unseen ones.

Overall, our contributions are threefold. (a) We propose C²SRT, which jointly models intra-category and inter-category semantic correlations for OV-MLR. (b) We introduce an ISR module that selects semantically meaningful local regions for each category, and an IST module that uses semantic priors to construct a category-adaptive correlation graph for transferring knowledge to unseen labels. (c) We extensively evaluate C²SRT on the NUS-WIDE and Open Images benchmarks under both zero-shot learning (ZSL) and generalized zero-shot learning (GZSL) settings, and provide in-depth ablations to analyze the contribution of each component.

II. RELATED WORK

A. Traditional MLR

Traditional multi-label methods often consider visual local features and label correlations. For local information, different regions of an image are typically evaluated based on their contribution to the target categories [24], [5], [25], [26]. For label correlations, semantic interactions between classes are achieved using graphs or other methods, as seen in [20], [4], [2], [6], [21], [1], which leverage co-occurrence or label similarity information to enable inter-category interactions. However, in the task of multi-label zero-shot learning, where unseen classes need to be recognized, an intuitive approach [27], [28] is to establish a connection between unseen and known classes by utilizing pretrained word embeddings such as GloVe [29] and lexical databases like WordNet. Recent studies, such as LESA [30] and BiAM [31], based on GloVe, capture both regional and global features for better multi-object recognition. While these methods facilitate information transfer between classes through language modalities and have shown some success, they struggle to address the challenges posed by open-vocabulary tasks.

B. Open-vocabulary MLR

In recent years, with the development of VLP models [32], [33], [34], open-vocabulary classification has emerged as an alternative to zero-shot prediction, achieving significant progress. Different OV settings in various application tasks, such as detection [35], [36], [37], segmentation [38], [39], [40] and scene understanding [41], [42], [43], have also been extensively explored. Leveraging billions of image-text pairs as training data, models like CLIP [44] and ALIGN [45] have achieved impressive performance in single-label zero-shot classification tasks. However, these methods are not fully adaptable to OV-MLR because VLP models are pretrained

for single-label classification by learning from one image-text pair, making them easily influenced by the image's dominant category. Consequently, recent works have begun exploring the use of VLP models for OV-MLR tasks. MKT [7] proposed a multi-modal knowledge transfer framework within VLP models, along with a dual-stream module for capturing both local and global features. However, MKT does not account for the correlations between labels in MLR, and its coarse, fixed handling of local features introduces noise. In this paper, we introduce a novel OV-MLR framework called the category-adaptive cross-modal semantic refinement and transfer (C²SRT), which adaptively handles intra-category local information and inter-category relationships.

III. METHOD

In this section, we first introduce the preliminary of open-vocabulary multi-label recognition and then describe the details of our proposed framework. Figure 3 illustrates the overall pipeline of our C²SRT framework.

A. Problem Setting

Let $\mathbf{x}_i \in \mathcal{X}$ denotes the i -th sample in the dataset, and $\mathbf{y}_i$ denotes the label present in this image. Particularly, $\mathbf{y}_i \in \mathcal{Y}^{\text{seen}}$ in the training set, and $\mathbf{y}_i \in \mathcal{Y}^{\text{seen}} \cup \mathcal{Y}^{\text{unseen}}$ in the test set. Here, $\mathcal{X}$, $\mathcal{Y}^{\text{seen}}$ and $\mathcal{Y}^{\text{unseen}}$ denote the image space of dataset, the set of seen labels, and unseen labels, respectively.

The goal of OV-MLR is to learn a classifier to identify all relevant labels in the given image, including seen labels and unseen labels. Specifically, two evaluation setups are widely used: (1) Zero-Shot Learning (ZSL): the classifier is exclusively evaluated by identifying unseen labels, which can be formulated as $f_{\text{ZSL}} : \mathcal{X} \rightarrow \mathcal{Y}^{\text{unseen}}$; (2) Generalized Zero-Shot Learning (GZSL): the classifier is tasked with identifying both seen and unseen labels, which can be formulated as $f_{\text{GZSL}} : \mathcal{X} \rightarrow \mathcal{Y}^{\text{seen}} \cup \mathcal{Y}^{\text{unseen}}$. Compared with the former, the latter is more challenging and realistic.

B. Vision Encoder with Knowledge Distillation

Given an input image $\mathbf{x}$, we first employ the vision transformer (ViT) [22] as the vision encoder to extract image features. Specifically, the image is divided into P non-overlapping patches and fed into the backbone along with a [CLS] token to generate the corresponding feature representations:

$$\mathcal{F}_p, f_G = \Phi_1(\mathbf{x}), \quad (1)$$

where $\mathbf{x} \in \mathcal{X}$ donates the input image, Φ_1 is the vision encoder, $\mathcal{F}_p \in \mathbb{R}^{P \times D}$ denotes the patch features, $f_G \in \mathbb{R}^D$ denotes the global feature derived from the [CLS] token, and D is the feature dimension of ViT.

The vision encoder is initialized from a vision-language pretraining model (i.e., CLIP [44]) and is fine-tuned during training. However, fine-tuning can cause the vision encoder to overfit the training data, thereby losing its ability to generalize to unseen categories. To address this, we adopt knowledge distillation during training to enhance the generalization capability of the vision encoder [46], [47]. The key to this process

is maintaining alignment between the global image feature extracted by the vision encoder Φ_1 and that from the original VLP, formulated as

$$\mathcal{L}_{\text{dist}} = \|f_G - f_G^{\text{CLIP}}\|_1, \quad (2)$$

where f_G^{CLIP} is the global image feature produced by the original pre-trained CLIP vision encoder.

C. Intra-category Semantic Refinement

Multi-label images inherently contain multiple objects from diverse semantic categories, which vary in size and are distributed across the entire image. Consequently, relying solely on the global features of the image often leads to the loss of critical visual cues and the introduction of noise. To address this limitation, in contrast to fixed-number patch representation used in current methods, we introduce the intra-category semantic refine (ISR) module that leverages the cross-modal knowledge of the VLP model to adaptively find a set of local discriminative regions that best represent the semantics of a target category.

To extract category-specific local features with better alignment, we leverage semantic guidance from the VLP text encoder. The textual feature $f_{\text{txt}}^{(c)}$ for a given category is obtained using the fixed VLP Text Encoder [48] Φ_T^{CLIP} :

$$f_{\text{txt}}^{(c)} = \Phi_T^{\text{CLIP}}(\text{prompt}_c), \quad (3)$$

where prompt_c represents the prompt corresponding to category c . To ensure both generalization and adaptability, the prompt for [CLS] is generated using an ensemble of common templates. For example, a common template is "A photo of [CLS]".

Then, ISR calculates the similarity between the i -th patch feature $f_p^{(i)}$ and text feature $f_{\text{txt}}^{(c)}$, denoted as $s_i^{(c)}$:

$$s_i^{(c)} = \text{Similarity}(f_p^{(i)}, f_{\text{txt}}^{(c)}). \quad (4)$$

The similarity of all patch features for category c , denoted as $S^{(c)} = [s_1^{(c)}, \dots, s_P^{(c)}]$, is passed through a softmax function to obtain $S^{(c)} = \text{SoftMax}(S^{(c)})$, representing the semantic matching scores for each local patch.

By sorting $S^{(c)}$ in descending order, we obtain the indices $k = [k_1, \dots, k_P]$, where $s_{k_i}^{(c)} \geq s_{k_j}^{(c)}$ for any $i \geq j$. Given a semantic threshold α , we select the semantic matching scores $S^{(c)}$ according to the order of indices k . If selecting up to k_n satisfies the condition $\sum_{i=1}^n s_{k_i}^{(c)} \geq \alpha$, then we consider the semantic alignment to be sufficient for the patches corresponding to $k_1, \dots, k_n$. We subsequently select the corresponding patch features and apply a pooling operation to compute the category-specific local features $f_L^{(c)}$, which are better aligned with the semantics of category c under its semantic guidance:

$$f_L^{(c)} = \text{Pooling}(f_p^{(k_1)}, \dots, f_p^{(k_n)}). \quad (5)$$

D. Inter-category Semantic Transfer

In traditional MLR, exploring inter-category correlations is proven useful, but it becomes quite challenging in OV-MLR due to the existence of unseen categories. To address this challenge, we propose the inter-category semantic transfer (IST)

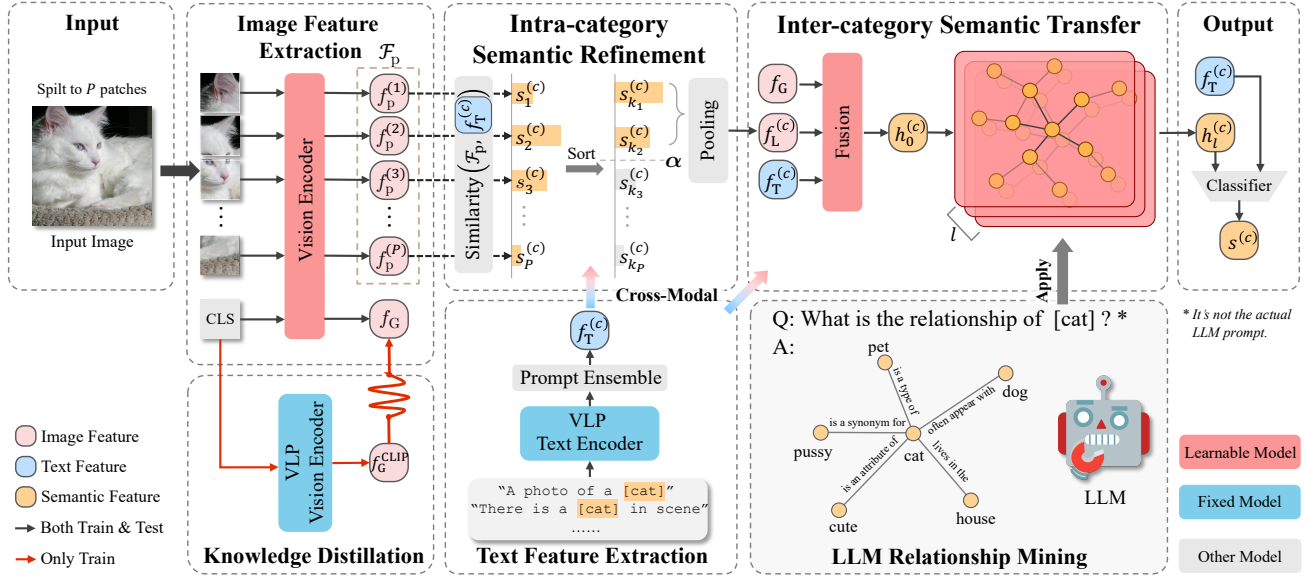


Fig. 3. The overall framework of our C²SRT framework. Our C²SRT employs a learnable vision encoder, which aligns features through knowledge distillation from a fixed VLP vision encoder, to extract image features. Simultaneously, a fixed VLP text encoder extracts ensemble-based textual features. The ISR module quantifies information by calculating the intra-category semantic similarity of local patch features, and adaptively focuses on local visual features using a threshold based on the total information. After the visual and textual features fusing, the multi-modal features are fed into the IST module, enabling adaptive inter-category knowledge transfer, with inter-category relationships derived from relation mining. Here, relationships are mined through an LLM.

module. It adaptively selects adjacent categories with rich contextual relationships for each category, thereby constructing an inter-category correlation graph that encapsulates flexible interactions for semantic transfer.

IST builds a sparse directed inter-category graph by connecting each category c to a small neighbor set $\mathcal{N}_c$ of the most related *seen* categories. The neighbor set $\mathcal{N}_c$ is obtained by a predefined category association metric. The directed edges indicate information flow from adjacent categories to the target category, enabling selective semantic transfer under a sparse graph topology.

- **Similarity-driven Inter-category Graph.** A simple way to construct $\mathcal{N}_c$ is through nearest-neighbor retrieval under text-embedding similarity (e.g., CLIP text features). In this way, we keep a fixed neighborhood size to ensure a sparse graph. We analyze this variant in Table IX.

- **LLM-driven Inter-category Graph.** In our default setting, we prompt a large language model (LLM) [49], [50], [51] to estimate association strengths between categories (see implementation details), and select a sparse neighbor set $\mathcal{N}_c$ of related seen categories for each c based on these scores.

Given the constructed sparse graph, we utilize graph attention networks (GAT) [17] to propagate information with adaptive edge weights, allowing the model to dynamically prioritize influential categories via learned attention coefficients.

First, we obtain the initial feature $h_0^{(c)}$ for category c :

$$h_0^{(c)} = \text{FFN}_{\text{in}}([f_{\text{img}}^{(c)} \parallel f_{\text{txt}}^{(c)}]), \quad (6)$$

where $f_{\text{img}}^{(c)} = (f_L^{(c)} + f_G)/2$ represents the image feature for category c , $f_{\text{txt}}^{(c)}$ is the text feature of category c , and $[\cdot \parallel \cdot]$ denotes vector concatenation. This results in the node feature

$h_0^{(c)} \in \mathbb{R}^{D_{\text{in}}}$ for category c , which is input into the first layer of the GAT.

Nodes are connected by edges, forming a graph, where $\mathcal{N}_i$ represents all nodes adjacent to node (category) i , and $j \in \mathcal{N}_i$ indicates that category i can receive information from category j . To obtain sufficient expressive power to transform the input features into higher-level features, a linear transformation is applied uniformly across all nodes, denoted by $W_a \in \mathbb{R}^{D_{\text{in}} \times D_{\text{hid}}}$, where D_{in} is the input node feature dimension, and D_{hid} is the hidden dimension. The attention coefficient between two nodes is then calculated by a shared attention mechanism with $a \in \mathbb{R}^{2 \cdot D_{\text{hid}} \times 1}$:

$$e_{ij} = a^\top \text{LeakyReLU}([W_a^\top h_0^{(i)} \parallel W_a^\top h_0^{(j)}]), \quad (7)$$

where e_{ij} quantifies the importance of node j to node i . For each node, only a subset of connected nodes $\mathcal{N}_i$ needs to be considered, specifically $j \in \mathcal{N}_i$. A SoftMax function is applied to these connected nodes to normalize the attention coefficients:

$$\alpha_{ij} = \text{SoftMax}_i(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}. \quad (8)$$

The output node features are weighted by the normalized attention coefficients α_{ij} to facilitate information transfer:

$$h_1^{(i)} = \sigma(\sum_{j \in \mathcal{N}_i} \alpha_{ij} W_{\text{out}}^\top h_0^{(j)}), \quad (9)$$

where $\sigma(\cdot)$ represents the activation function, and $h_1^{(i)}$ is the node feature of category i output in the first GAT layer. $W_{\text{out}} \in \mathbb{R}^{D_{\text{in}} \times D_{\text{hid}}}$ denotes the learnable output transformation used in message aggregation to map neighbor features from the input space to the hidden space before weighted summation.

After stacking l layers of GAT, the output $h_l^{(i)}$ is used for category prediction. To achieve better inter-category interaction performance, we implement GATv2 [52], which introduces a multi-head mechanism. We use a shallow graph encoder in IST (default $l = 2$ layers with multi-head attention) since the inter-class graph is sparse and built from a small set of semantic neighbors per class; 1–2 hops of message passing typically suffice to propagate the most relevant knowledge. Increasing the depth tends to introduce less relevant information and may cause over-smoothing and optimization instability in deep GNNs [53], [54], [55]. We therefore adopt a shallow configuration as a robust trade-off between effective transfer and stable training.

Prediction. Following previous works, the prediction score is computed by the similarity between the output feature and the corresponding text feature of category i :

$$\hat{y}_i = \text{Similarity}(h_l^{(i)}, f_{\text{txt}}^{(i)}), \quad (10)$$

where $\hat{y}_i$ is the model's prediction score for the i -th category, and $\text{Similarity}(\cdot, \cdot)$ denotes the cosine similarity function as employed in CLIP.

Optimization. In this work, we utilize the ranking loss as classification loss, formulated as

$$\mathcal{L}_{\text{cls}} = \sum_k \sum_{p \in \mathbf{y}_{\text{pos}}^{(k)}, n \notin \mathbf{y}_{\text{pos}}^{(k)}} \max(\hat{y}_p^{(k)} - \hat{y}_n^{(k)} + 1, 0), \quad (11)$$

where $\mathbf{y}_{\text{pos}}^{(k)} = \{j : \mathbf{y}_j^{(k)} = 1\}$ represents the positive labels in the ground truth for image k , and $\mathbf{y}_j^{(k)}$ indicates the label of category j for image k . The indices p and n denote the positive and negative labels in the ground truth of the image k , respectively, while $\hat{y}_p^{(k)}$ and $\hat{y}_n^{(k)}$ are the corresponding prediction scores. The goal is to ensure that the scores of positive labels are ranked higher than those of negative labels by a minimum margin of 1.

The final model loss is computed as the sum of the classification loss and the distillation loss with its weight γ :

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \gamma \mathcal{L}_{\text{dist}}. \quad (12)$$

IV. EXPERIMENTS

A. Experiment Setup

Dataset. We validate the superiority of our model using two widely recognized benchmarks. **NUS-WIDE** [60] is a comprehensive web dataset. It comprises a training set of 161,789 images and a testing set of 107,859 images. Following the LESA setting, we treat 81 human-verified labels as unseen labels, and 925 labels generated user tags as seen labels. The **Open Images (v4)** [61] dataset includes 9,011,219 images for training, 41,620 images for validation, and 125,436 images for testing. As per LESA [30], we designate 7,186 labels with more than 100 images in the training set as seen labels, and the 400 most frequent test labels that do not appear in the training set as unseen labels.

Metrics. Following previous works, we adapt the precision (P), recall (R), and F1 score (F1) to evaluate models. Though balancing the trade-off between precision and recall, the F1

score offers a comprehensive measure of overall performance. Additionally, we also introduce the metric of mean Average Precision (mAP) over all categories.

B. Implementation Details

We utilize pre-trained CLIP [44] as our VLP model, with ViT-B/16 serving as the vision encoder and a Transformer as the text encoder. ViT-B/16 also functions as a student model for distillation. Images are pre-processed to 224×224 pixels, with the Vision Encoder dividing each image into $14 \times 14 = 196$ patches. In ISR, the maximum number of patches N is set to 32, with an information threshold α of 0.5.

During training, we employ the AdamW optimizer with the learning rate of 1×10^{-3} and a weight decay of 5×10^{-3} . For the NUS-WIDE dataset, the model was trained for 20 epochs with a batch size of 64. IST selects 16 related categories, and the GAT model comprises 2 layers. For the Open Images dataset, we train for 8 epochs, with IST selecting 4 related categories. All other configurations remain consistent with those used for the NUS-WIDE dataset.

LLM Relation Mining Details. We use the OpenAI API with model id GPT-4o-2024-08-06. We use deterministic decoding (temperature=0, top_p=1) and enforce a strict output format to reduce randomness. If parsing fails, we retry a fixed number of times using the same prompt and decoding parameters.

The prompt template is constructed as follows. Given a target category (**New Category**) and a candidate list of seen categories (**List of Seen Categories**), we ask the LLM to identify directly related categories. Relationship types are defined as: (1) *Synonymy/Similarity*; (2) *Is-a/Hypernym*; (3) *Functional Relationship*; (4) *Co-occurrence*; (5) *Part-Whole Relationship*. For each related category, the LLM outputs (i) relationship type, (ii) association strength (High/Medium/Low), and (iii) a brief explanation. The output is constrained to only include category names from the provided candidate list.

We perform post-processing for graph construction. We map association strength to numerical scores (High=3, Medium=2, Low=1, not mentioned=0). Each category is queried multiple times and the scores are averaged. We then select top- N related categories by descending average score to construct a sparse directed graph for GAT message passing. The one-time cost for constructing a class-relation graph (about 1000 classes) is approximately \$2.25.

C. Comparisons with State-of-the-art Methods

We compare our model with ZSL models and OV models. The results for ZSL and GZSL are shown in Table I. Consistent with previous methods, we calculate the F1 scores by selecting the top-3 and top-5 categories for NUS-WIDE, and the top-10 and top-20 categories for Open Images.

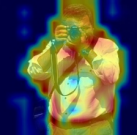
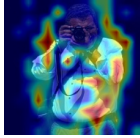
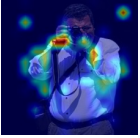
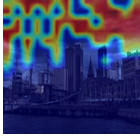
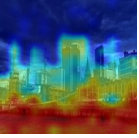
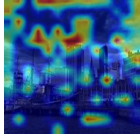
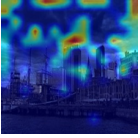
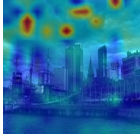
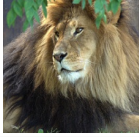
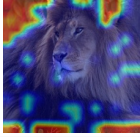
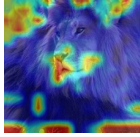
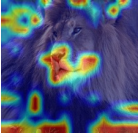
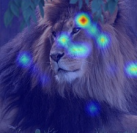
On the NUS-WIDE dataset, our model improves over MKT in both ZSL and GZSL. In GZSL, mAP increases from 18.3% to 19.6% (7.1% relative), and F1 improves by 5.0% at $K = 3$ and 4.3% at $K = 5$. In ZSL, mAP improves by 1.6% relative, and F1 improves by 1.5% at $K = 3$ and 3.5% at $K = 5$.

On the Open Images dataset, our model also outperforms MKT. In ZSL, F1 improves by 6.1% relatively at $K = 10$

TABLE I
COMPARISON (%) WITH STATE-OF-THE-ART WITH ZS-MLR AND OV-MLR METHODS ON NUS-WIDE AND OPEN IMAGES DATASETS UNDER THE ZSL AND GZSL SETTINGS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Setting	Task	NUS-WIDE							Open Images						
			P	K=3 R	F1	P	K=5 R	F1	mAP	P	K=10 R	F1	P	K=20 R	F1	mAP
Fast0Tag [56]	ZS	ZSL	22.6	36.2	27.8	18.2	48.4	26.4	15.1	-	-	-	-	-	-	-
		GZSL	18.8	8.3	11.5	15.9	11.7	13.5	3.7	-	-	-	-	-	-	-
LESA [30]		ZSL	25.7	41.1	31.6	19.7	52.5	28.7	19.4	0.7	25.6	1.4	0.5	37.4	1.0	41.7
		GZSL	23.6	10.4	14.4	19.8	14.6	16.8	5.6	16.2	18.9	17.4	10.2	23.9	14.3	45.4
(ML) ² -Enc[57]		ZSL	-	-	32.8	-	-	32.3	29.4	-	-	7.5	-	-	6.5	65.7
		GZSL	-	-	15.8	-	-	19.2	10.2	-	-	27.6	-	-	24.1	79.9
ZS-SDL [58]		ZSL	24.2	41.3	30.5	18.8	53.4	27.8	25.9	6.1	47.0	10.7	4.4	68.1	8.3	62.9
		GZSL	27.7	13.9	18.5	23.0	19.3	21.0	12.1	25.3	40.8	37.8	23.6	54.5	32.9	75.3
BiAM [31]		ZSL	26.6	42.5	32.7	20.5	54.6	29.8	25.9	3.9	30.7	7.0	2.7	41.9	5.5	65.6
		GZSL	25.2	11.1	15.4	21.6	15.9	18.2	9.4	13.8	15.9	14.8	9.7	22.3	14.8	81.7
CLIP [44]	OV	ZSL	27.0	33.5	29.9	21.2	43.8	28.5	33.7	10.8	84.0	19.1	5.9	92.1	11.1	66.2
		GZSL	31.4	13.7	19.0	26.0	18.9	21.9	16.5	37.5	43.3	40.2	25.4	58.7	35.4	77.5
Tag-CLIP [59]		ZSL	26.4	42.1	32.4	19.3	51.5	28.1	38.5	-	-	-	-	-	-	-
		GZSL	24.3	10.7	14.9	19.6	14.4	16.0	13.1	-	-	-	-	-	-	-
MKT [7]		ZSL	27.7	44.3	34.1	21.3	57.0	31.1	37.6	11.1	86.8	19.7	6.1	94.7	11.4	68.1
		GZSL	35.9	16.8	22.0	29.9	22.0	25.4	18.3	37.8	43.6	40.5	25.4	58.5	35.4	81.4
Ours		ZSL	28.1	45.0	34.6	22.1	59.0	32.2	39.2	11.9	87.0	20.9	6.6	94.3	12.4	69.0
		GZSL	37.7	16.6	23.1	31.3	23.0	26.5	19.6	38.2	44.1	40.9	25.2	60.0	35.5	82.1

TABLE II
SEVERAL EXAMPLES OF CLASS ACTIVATION MAPS (CAM) FROM THE IMAGE ENCODER OF CLIP [44], MKT [7], AND OUR PROPOSED METHOD. SINCE THE PATCH FEATURES FROM THE FINAL LAYER OF CLIP’S IMAGE ENCODER ARE NOT UTILIZED, THE OUTPUT FROM THE PENULTIMATE LAYER IS EMPLOYED INSTEAD.

Origin	Label	CLIP		MKT		Ours	
		positive	negative	positive	negative	positive	negative
	(p) Camera (n) Pets						
	(p) Sky (n) Dog						
	(p) Lion (n) Phone						

(19.7% to 20.9%) and by 8.8% at $K = 20$, and mAP increases by 1.3%. In GZSL, our model improves over MKT by 1.2% relatively in F1 score at $K = 10$ and by 0.9% in mAP.

Overall, these results demonstrate that our method achieves consistent gains over MKT across datasets and metrics.

Visualization of Class Activation Map. Table II shows the class activation mapping (CAM) [66] of the CLIP, MKT, and our method. It can be observed that for the correct categories, CLIP, MKT, and our method all focus on the correct regions. However, for incorrect categories, both CLIP and MKT activate a large number of incorrect regions, which can lead to

erroneous predictions. In contrast, when analyzing incorrect categories, our model focuses less on incorrect regions, thereby achieving the effect of suppressing incorrect categories.

Comparison with Multi-Modal LLMs (MLLMs). Beyond methods specifically designed for OV-MLR, we also compare our framework with representative MLLMs, which achieve recognition via language generation. Considering the substantially larger model sizes and inference costs of MLLMs, we evaluate them on a randomly sampled subset of 1000 images from NUS-WIDE under a unified “prompt-only open-vocabulary multi-label prediction” protocol. As shown in Ta-

TABLE III
COMPARISON (%) WITH REPRESENTATIVE MLLMs ON A RANDOMLY SAMPLED SUBSET OF 1000 NUS-WIDE IMAGES UNDER THE “PROMPT-ONLY OPEN-VOCABULARY MULTI-LABEL PREDICTION” PROTOCOL.

Method	ZSL			GZSL		
	P	R	F1	P	R	F1
LLaVA1.5 [62]	32.6	20.3	25.1	43.8	8.4	14.1
LLaVA1.6 [63]	35.6	17.3	23.3	48.0	7.4	12.8
LLaMA3.2V [64]	28.8	55.5	37.9	42.1	25.6	31.8
Qwen2.5VL [65]	91.7	51.5	66.0	95.1	23.0	37.0
Ours	79.2	61.1	69.0	65.2	36.6	46.9

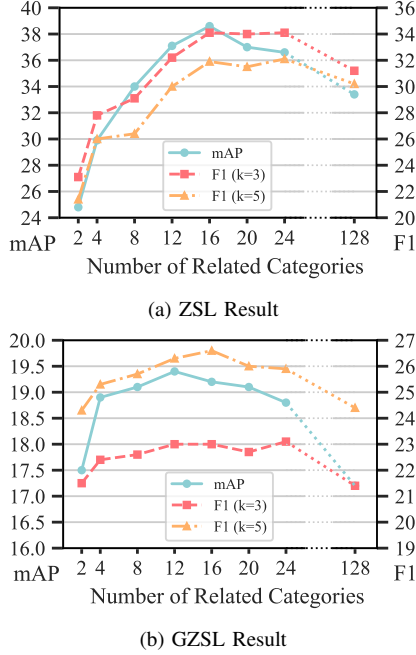


Fig. 4. Effect of varying numbers of related categories in the IST module for (a) zero-shot learning (ZSL) and (b) generalized zero-shot learning (GZSL) tasks on the NUS-WIDE dataset.

ble III, our model (88M parameters) achieves stronger overall F1 compared with these MLLMs, with particularly larger gains under GZSL. While Qwen2.5-VL achieves higher precision, our recall is consistently higher, leading to better overall F1.

D. Ablation Study

Effect of Distillation, ISR and IST. To evaluate the impact of feature alignment through knowledge distillation, the ISR module, and the IST module, we conduct ablation experiments on NUS-WIDE under consistent configurations and report the mean $\pm$ std results over 3 runs in Table IV. As can be observed, introducing distillation improves ZSL and GZSL performance, and adding ISR further improves ZSL metrics by leveraging category-adaptive local evidence. Adding IST improves top- K F1, especially under GZSL, by transferring complementary semantics across categories. Notably, C²SRT achieves the best F1@3/F1@5 in both ZSL and GZSL, while Dist+IST yields a slightly higher GZSL mAP, indicating that IST can be particularly beneficial for ranking quality under GZSL.

TABLE IV
IMPACT OF KNOWLEDGE DISTILLATION (DIST), INTRA-CATEGORY SEMANTIC REFINEMENT (ISR), AND INTER-CATEGORY SEMANTIC TRANSFER (IST) ON NUS-WIDE. RESULTS ARE REPORTED AS MEAN $\pm$ STD OVER 3 RUNS. “X” DENOTES THE ABSENCE OF THE MODULE. “✓” INDICATES THE PRESENCE OF THE MODULE.

Module	Dist	ISR	IST	Task	mAP	F1	
						K=3	K=5
X	X	X		ZSL	32.5 $\pm$ 0.5	29.1 $\pm$ 0.6	26.8 $\pm$ 0.6
				GZSL	16.9 $\pm$ 0.6	20.1 $\pm$ 1.0	25.2 $\pm$ 1.1
✓	✓	X	X	ZSL	37.5 $\pm$ 0.2	32.5 $\pm$ 0.2	29.7 $\pm$ 1.9
				GZSL	17.8 $\pm$ 0.5	21.3 $\pm$ 0.8	25.6 $\pm$ 0.7
✓	✓	✓	X	ZSL	39.0 $\pm$ 1.0	33.3 $\pm$ 0.8	32.1 $\pm$ 0.8
				GZSL	18.4 $\pm$ 0.6	22.5 $\pm$ 1.4	24.8 $\pm$ 0.9
✓	✓	X	✓	ZSL	37.6 $\pm$ 0.6	34.3 $\pm$ 0.8	31.0 $\pm$ 0.9
				GZSL	19.1$\pm$0.7	23.0 $\pm$ 1.0	26.1 $\pm$ 0.3
✓	✓	✓	✓	ZSL	39.4$\pm$0.6	34.4$\pm$0.3	32.5$\pm$0.6
				GZSL	18.5 $\pm$ 1.1	23.2$\pm$0.6	26.2$\pm$0.4

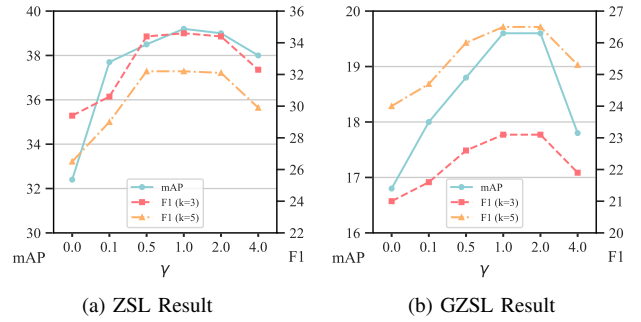


Fig. 5. Effect of the distillation-weight hyper-parameter γ in Eq. (12) for (a) ZSL and (b) GZSL tasks on the NUS-WIDE dataset.

Multi-seed Statistics and Significance Tests. For re-runnable baselines under a unified training/evaluation protocol (same data splits, image preprocessing, epochs, and evaluation script), we repeat training three times and report mean $\pm$ std in Table V. Our method shows consistent performance gains. Here, CLIP/MKT/C²SRT are re-trained by us under the unified protocol for fair comparison. In contrast, Table I primarily summarizes results reported in the original papers for other methods.

Confidence Intervals and Significance Tests. For key ablations (Dist/ISR/IST), we additionally report 95% confidence intervals (t-distribution, df=2) and two-sided paired t-tests comparing the full model (Dist+ISR+IST) with each ablation, as summarized in Tables VI and VII.

Sensitivity to Parameter γ . We further conduct experiments to analyze the impact of the distillation weight γ in our loss function. As shown in Figure 5, an intermediate γ (around 1.0) yields more stable performance, while too small or too large values may under-emphasize distillation or over-regularize the student, respectively. We observe consistent trends under an alternative CLIP backbone (RN50), as shown in Table VIII.

Analysis of IST Module. We investigated the effect of replacing the adjacent categories extracted by LLM in IST module with either randomly selected categories or textual similarity-based categories, as detailed in Table IX. The results

TABLE V
MEAN $\pm$ STD OVER 3 RUNS FOR RE-RUNNABLE BASELINES UNDER A UNIFIED TRAINING/EVALUATION PROTOCOL.

Method	ZSL			GZSL		
	F1@3	F1@5	mAP	F1@3	F1@5	mAP
CLIP	29.6 $\pm$ 0.3	28.8 $\pm$ 0.3	33.3 $\pm$ 1.2	18.4 $\pm$ 0.6	22.0 $\pm$ 0.6	16.0 $\pm$ 1.2
MKT	33.7 $\pm$ 1.2	31.2 $\pm$ 0.2	37.8 $\pm$ 0.4	22.4 $\pm$ 1.4	25.7 $\pm$ 0.2	19.0 $\pm$ 0.8
Ours	34.7$\pm$0.3	32.0$\pm$0.5	39.5$\pm$0.3	24.2$\pm$1.0	26.4$\pm$0.8	19.2$\pm$0.5

TABLE VI
MEAN $\pm$ STD AND 95% CONFIDENCE INTERVALS (CI95) FOR ABLATIONS UNDER ZSL AND GZSL.

Module			Task	mAP	F1	
Dist	ISR	IST			K=3	K=5
$\times$	$\times$	$\times$	ZSL	32.5 $\pm$ 0.5 [31.3,33.8]	29.1 $\pm$ 0.6 [27.5,30.7]	26.8 $\pm$ 0.6 [25.2,28.4]
			GZSL	16.9 $\pm$ 0.6 [15.5,18.3]	20.1 $\pm$ 1.0 [17.7,22.5]	25.2 $\pm$ 1.1 [22.5,27.9]
$\checkmark$	$\times$	$\times$	ZSL	37.5 $\pm$ 0.2 [37.0,38.0]	32.5 $\pm$ 0.2 [32.0,33.0]	29.7 $\pm$ 1.9 [25.1,34.3]
			GZSL	17.8 $\pm$ 0.5 [16.5,19.1]	21.3 $\pm$ 0.8 [19.4,23.2]	25.6 $\pm$ 0.7 [24.0,27.2]
$\checkmark$	$\checkmark$	$\times$	ZSL	39.0 $\pm$ 1.0 [36.5,41.5]	33.3 $\pm$ 0.8 [31.3,35.3]	32.1 $\pm$ 0.8 [30.2,33.9]
			GZSL	18.4 $\pm$ 0.6 [16.9,20.0]	22.5 $\pm$ 1.4 [19.1,25.9]	24.8 $\pm$ 0.9 [22.6,27.0]
$\checkmark$	$\times$	$\checkmark$	ZSL	37.6 $\pm$ 0.6 [36.0,39.2]	34.3 $\pm$ 0.8 [32.3,36.3]	31.0 $\pm$ 0.9 [28.7,33.3]
			GZSL	19.1 $\pm$ 0.7 [17.4,20.9]	23.0 $\pm$ 1.0 [20.6,25.4]	26.1 $\pm$ 0.3 [25.4,26.7]
$\checkmark$	$\checkmark$	$\checkmark$	ZSL	39.4 $\pm$ 0.6 [38.0,40.8]	34.4 $\pm$ 0.3 [33.6,35.2]	32.5 $\pm$ 0.6 [31.0,34.0]
			GZSL	18.5 $\pm$ 1.1 [15.8,21.2]	23.2 $\pm$ 0.6 [21.8,24.6]	26.2 $\pm$ 0.4 [25.3,27.0]

Note: Each cell shows mean $\pm$ std (line 1) and CI95 (line 2). CI95 is computed with a t-distribution over 3 runs (df=2).

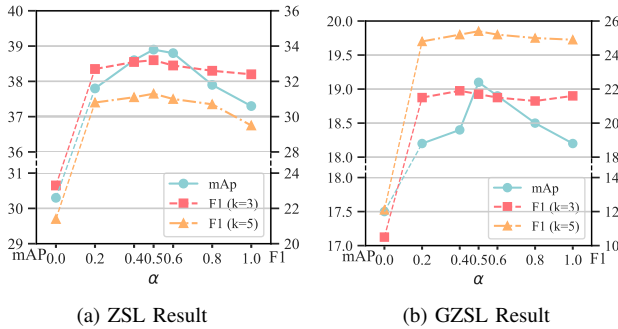


Fig. 6. Effect of hyper-parameter α in the ISR module for (a) zero-shot learning (ZSL) and (b) generalized zero-shot learning (GZSL) tasks on the NUS-WIDE dataset. Note that $\alpha = 0.0$ indicates the absence of local features.

demonstrate that substituting adjacent categories leads to a decline in performance for both ZSL and GZSL. However, the performance degradation in GZSL is smaller due to the inclusion of seen labels, underscoring the model's robustness. In contrast, ZSL, which considers solely on unseen labels, resulting in a substantial performance declines because erroneous

TABLE VII
SIGNIFICANCE TESTS (TWO-SIDED PAIRED T-TEST) COMPARING THE FULL MODEL AGAINST EACH ABLATION UNDER ZSL AND GZSL.

Module			Task	mAP	F1	
Dist	ISR	IST			K=3	K=5
$\times$	$\times$	$\times$	ZSL	+6.8 (0.008)	+5.2 (0.001)	+5.7 (0.013)
			GZSL	+1.6 (0.208)	+3.1 (0.024)	+1.0 (0.359)
$\checkmark$	$\times$	$\times$	ZSL	+1.9 (0.014)	+1.9 (0.004)	+2.8 (0.178)
			GZSL	+0.7 (0.192)	+1.9 (0.019)	+0.6 (0.356)
$\checkmark$	$\checkmark$	$\times$	ZSL	+0.4 (0.699)	+1.0 (0.106)	+0.4 (0.526)
			GZSL	+0.1 (0.843)	+0.7 (0.606)	+1.4 (0.050)
$\checkmark$	$\times$	$\checkmark$	ZSL	+1.8 (0.092)	+0.1 (0.891)	+1.5 (0.228)
			GZSL	-0.6 (0.605)	+0.2 (0.840)	+0.1 (0.775)

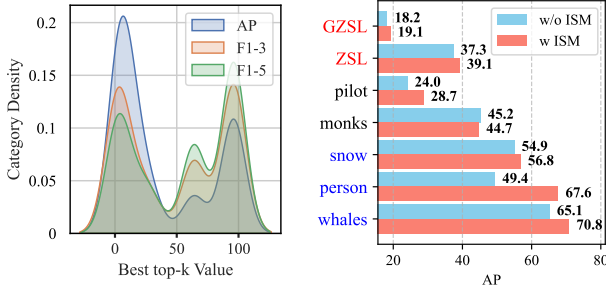
Note: Each cell shows Δ (line 1) and p-value (line 2). Δ = Full – Variant. Two-sided paired t-tests are conducted over 3 runs (df=2).

TABLE VIII
SENSITIVITY TO DISTILLATION WEIGHT γ IN EQ. (12) ON NUS-WIDE WITH CLIP RN50 (OTHER HYPER-PARAMETERS FIXED).

γ	ZSL			GZSL		
	F1@3	F1@5	mAP	F1@3	F1@5	mAP
0	28.2	27.7	30.1	20.4	24.6	12.7
1	32.8	30.1	37.2	22.3	25.0	16.9
2	31.7	30.1	37.9	22.4	24.5	17.0

information transfer cannot be effectively mitigated during training. Furthermore, similarity-based replacement outperforms random selection because semantic similarity inherently captures a certain degree of association, whereas random selection introduces more noise. Nonetheless, similarity-based methods still cannot fully capture the complex inter-category relationships.

Figure 4 explores the impact of the number of adjacent categories on model performance. The results indicate that a moderate number of adjacent categories yields the best performance for both ZSL and GZSL. Specifically, having too few adjacent categories significantly impairs ZSL performance due to the impact of absent inter-category information transfer. But in GZSL, the presence of seen categories during training reduces, thereby having a relatively limited effect on performance. Conversely, an excessive number of related categories



(a) Distribution of Optimal Top-k (b) Effect of the ISR module on AP

Fig. 7. Ablation analysis of ISR on the NUS-WIDE dataset: (a) Distribution of top-k patch selections per category for optimal mAP and F1 scores. (b) Impact of the ISR module on AP in GZSL and ZSL settings, with blue indicating unseen categories.

TABLE IX

ABLATION STUDY ON DIFFERENT INTER-CATEGORY RELATIONSHIPS IN THE IST. “RANDOM” REFERS TO RANDOMLY GENERATED RELATIONSHIPS. “SIMILARITY” REFERS TO RELATIONSHIPS DERIVED FROM TEXT EMBEDDING SIMILARITIES. “LLM” REFERS TO RELATIONSHIPS MINED USING LLM.

Relation	Task	mAP	F1	
			K=3	K=5
Random	ZSL	33.4	30.2	27.8
	GZSL	18.1	21.3	24.3
Similarity	ZSL	35.5	31.2	28.3
	GZSL	19.1	22.3	24.9
LLM	ZSL	39.2	34.1	31.9
	GZSL	19.2	23.0	26.6

leads to performance degradation in both ZSL and GZSL, as not all categories contribute positively to recognition and the increased complexity hinders training. This trend also holds under CLIP RN50 on NUS-WIDE, as shown in Table X.

Analysis of ISR Module. Figure 6 demonstrates the impact of varying α values in the ISR on the NUS-WIDE dataset. A larger α indicates the selection of more local feature information, whereas a smaller α implies fewer local features. Specifically, $\alpha = 1$ corresponds to utilizing all local features, and $\alpha = 0$ denotes the exclusion of local features. The results clearly show that omitting local features significantly degrades performance, underscoring the critical role of local feature integration. As α increases, performance initially improves due to the beneficial contribution of appropriate local information to recognition. However, beyond a certain point (about 0.5), particularly at $\alpha = 1$, the introduction of excessive local features introduces noise, leading to a decline in performance.

We conduct further ablation experiments across backbones and datasets to analyze the sensitivity of α . Besides the default setting (CLIP ViT-B/16 on NUS-WIDE), we further examine sensitivity trends under (i) an alternative CLIP backbone (RN50) on NUS-WIDE, and (ii) the Open Images dataset with ViT-B/16 (following our Open Images training setup). We confirm that taking α around 0.5 remains a stable choice, as shown in Tables XI and XII.

Figure 7(a) illustrates that different categories achieve optimal performance with varying numbers of local features. This

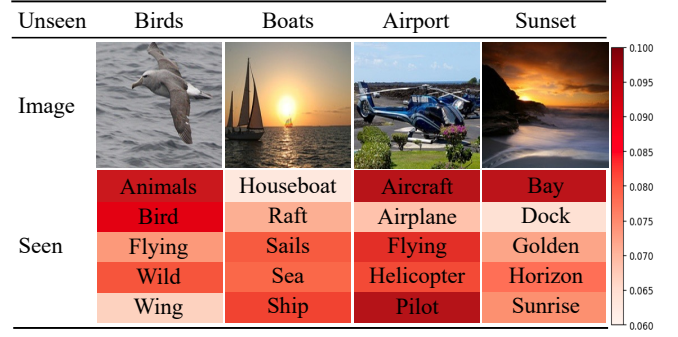


Fig. 8. Visualization of the relational weights heatmap between unseen labels and their top-5 related seen categories.

TABLE X

SENSITIVITY TO THE NUMBER OF RELATED CATEGORIES (#related) ON NUS-WIDE WITH CLIP RN50 (FIXED $\alpha = 0.5$).

#related	ZSL			GZSL		
	F1@3	F1@5	mAP	F1@3	F1@5	mAP
4	30.8	29.3	36.4	22.1	24.2	16.6
16	32.8	30.1	37.2	22.3	25.0	16.9
32	28.4	25.6	33.9	18.8	21.1	14.6

TABLE XI

SENSITIVITY TO ISR THRESHOLD α ON NUS-WIDE WITH CLIP RN50 (OTHER HYPER-PARAMETERS FIXED).

α	ZSL			GZSL		
	F1@3	F1@5	mAP	F1@3	F1@5	mAP
0.0	28.5	24.9	29.6	16.8	19.8	12.3
0.2	30.2	26.5	34.3	19.4	22.7	16.7
0.5	32.8	30.1	37.2	22.3	25.0	16.9
0.8	30.4	28.8	38.1	20.6	22.3	15.2
1.0	31.5	29.1	36.2	18.3	19.4	12.6

TABLE XII

SENSITIVITY TO ISR THRESHOLD α ON OPEN IMAGES WITH CLIP ViT-B/16 (OTHER HYPER-PARAMETERS FIXED).

α	ZSL			GZSL		
	F1@10	F1@20	mAP	F1@10	F1@20	mAP
0.2	18.8	10.0	66.9	39.3	34.2	80.8
0.5	20.9	12.4	69.0	40.6	35.5	82.1
0.8	16.4	9.8	64.3	37.7	34.1	78.2

variability is attributed to differences in category appearance and discriminative region, highlighting the necessity for adaptive local feature refinement. Figure 7(b) further reveals that incorporating the ISR module, which leverages semantically guided adaptive local features, enhances the mAP metrics for both ZSL and GZSL. Notably, certain categories experience significant performance improvements, demonstrating the effectiveness of the ISR in adapting to category-specific feature requirements.

E. Efficiency and Resource Profile

Inference-side Profiling. To analyze the model’s inference efficiency, we report parameter count, FLOPs, inference latency,

TABLE XIII
INFERENCE-SIDE EFFICIENCY PROFILING. FLOPS ARE PER FORWARD PASS. *infer. (ms)* REPORTS MEAN / P50 LATENCY. *memory* REPORTS PEAK GPU MEMORY (MB).

Model	Params	FLOPs	infer. (ms)	memory
CLIP (backbone)	86.193M	33.869G	4.47 / 4.43	371.5
MKT	86.193M	34.056G	4.71 / 4.55	397.6
C ² SRT w/o ISR&IST	86.193M	34.208G	4.64 / 4.54	370.1
C ² SRT w/o IST	86.193M	34.428G	4.68 / 4.58	373.6
C ² SRT (full)	88.296M	38.332G	6.46 / 6.36	442.2

and peak GPU memory for representative variants under a unified setup (image_size=224, batch_size=1, measured on an RTX 4090). The results are summarized in Table XIII.

As can be observed, the difference between C²SRT w/o ISR&IST and C²SRT w/o IST reflects the incremental overhead introduced by ISR, and the difference between C²SRT w/o IST and the full C²SRT reflects the additional cost of IST (graph interactions via GAT). The comparison indicates that IST consumes more inference time and memory than ISR. We note that the LLM is only used offline to construct the inter-class relationship graph (cached and reused), and no LLM is needed during inference.

Training-side Time and Memory. Under the default training setup (CLIP ViT-B/16, image_size=224, training with 20 epochs and batch_size=64), the model training on the NUS-WIDE dataset takes about 20 minutes on 2×RTX 4090, and the peak GPU memory per card is about 25 GB.

F. Qualitative Analysis

Visualization of Category Relationships. As shown in Figure 8, the IST module is capable of adaptively transferring information from seen categories when recognizing unseen categories. It can be observed that the top-5 correlation coefficients of the unseen categories are closely related to the seen categories within the images.

Evaluation of Open-Vocabulary Recognition. To evaluate the open-vocabulary capabilities, we select novel images and categories that are absent from the evaluation dataset and rarer as well as more challenging. The results shown in Figure 9 indicate that MKT’s classification performance is inferior to that of CLIP, potentially due to bias introduced during training. In contrast, our method shows superior recognition ability in this open-vocabulary setting, effectively utilizing information from seen categories to enhance the recognition of novel ones, thus demonstrating significant potential.

Failure Case Analysis. To understand the proposed C²SRT method more comprehensively, Figure 10 presents three representative failure cases on NUS-WIDE. (a) A landmark with a tower-like outline resembles the visual prototype of “university”, causing confusion between tower and similar building labels for our method. (b) Holiday decorations activate the holiday/holidays/Christmas cluster incorrectly for our method, producing within-cluster false positives and lowering the rank of the true identity/attribute labels. (c) The cue for mirror is vague and partly occluded, leaving ISR with little usable evidence. Our model thus falls back to coarse

		
CLIP	MKT	Ours
dodo	birds	birds
animal	dodo	dodo
person	animal	animal
birds	coral	tree
toy	leaf	rocks
flowers	moth orchid	flowers
cypripedium japonicum	flowers	cypripedium japonicum
plants	cypripedium japonicum	leaf
moth orchid	plants	plants
garden	garden	garden
person	person	awkward
awkward	tattoo	person
angry	rainbow	angry
animal	window	tree
dancing	cat	house

Fig. 9. The top-5 category prediction results of each model in the open-vocabulary setting. Green indicates positive, red indicates negative. The yellow background indicates a novel category in the open-vocabulary setting, otherwise, it is an unseen category from the dataset.

			
PD	GT	PD	GT
university	India	holiday	red
cityscape	photographer	holidays	film
sky	clouds	Christmas	beaty
cathedral	sky	colorful	actor
Sydney	tower		
		room	lights
		panorama	mirror
		dance	stars
		playing	actor
		shop	delate
(a)		(b)	(c)

Fig. 10. Failure-case examples on NUS-WIDE. PD denotes the model prediction and GT denotes the ground truth. Green indicates true positives (TP), red indicates false positives (FP), and yellow background indicates false negatives (FN).

scene labels (e.g., room), causing missed recognition of mirror and related labels (false negatives). Beyond visual factors, some apparent errors can also stem from label ambiguity or incomplete annotations (e.g., overlapping concepts, synonyms, or missing positive tags), where a visually plausible prediction is counted as a false positive by evaluation. These cases also match our sensitivity analysis: a large α in ISR can admit noisy patches, and an overly large related set in IST can propagate irrelevant semantics.

V. CONCLUSION

We propose C²SRT, a category-adaptive framework for OV-MLR that combines intra-category semantic refinement (ISR) and inter-category semantic transfer (IST). The ISR module extracts category-adaptive local evidence, while the IST module transfers semantics from related seen categories to unseen ones via a category-adaptive correlation graph. Extensive experiments on NUS-WIDE and Open Images benchmarks show that our method achieves consistent improvements over current leading methods.

While this work focuses on image-based OV-MLR, the core ideas of ISR/IST naturally extend to video/temporal settings. In videos, ISR can generalize patch selection from single frames to temporal tubes/tracklets, and exploit cross-frame

consistency to suppress noisy local evidence and reduce background shortcuts. IST can further incorporate temporal context by combining category-graph propagation with temporal attention, enabling joint transfer across categories and across time. However, video datasets often exhibit more severe label noise and long-tail distributions, which may amplify errors in LLM-based relation mining and graph propagation. Computation and storage costs also increase, motivating stronger efficiency strategies (e.g., offline relation caching, sparse temporal sampling, and lightweight temporal aggregation).

REFERENCES

- [1] J. Zhong, R. Shang, F. Zhao, W. Zhang, and S. Xu, "Negative label and noise information guided disambiguation for partial multi-label learning," *IEEE Transactions on Multimedia*, vol. 26, pp. 9920–9935, 2024.
- [2] X. Zhu, J. Liu, W. Liu, J. Ge, B. Liu, and J. Cao, "Scene-aware label graph learning for multi-label image classification," in *Int. Conf. Comput. Vis.*, 2023, pp. 1473–1482.
- [3] T. Pu, Q. Lao, H. Wu, T. Chen, L. Tian, J. Liu, and L. Lin, "Category-adaptive label discovery and noise rejection for multi-label recognition with partial positive labels," *IEEE Trans. Multimedia*, vol. 26, pp. 9591–9602, 2024.
- [4] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 5177–5186.
- [5] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Int. Conf. Comput. Vis.*, 2017, pp. 464–472.
- [6] J. Yuan, S. Chen, Y. Zhang, Z. Shi, X. Geng, J. Fan, and Y. Rui, "Graph attention transformer network for multi-label image classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 4, pp. 150:1–150:16, 2023. [Online]. Available: <https://doi.org/10.1145/3578518>
- [7] S. He, T. Guo, T. Dai, R. Qiao, X. Shu, B. Ren, and S. Xia, "Open-vocabulary multi-label classification via multi-modal knowledge transfer," in *AAAI Conf. Artif. Intell.*, B. Williams, Y. Chen, and J. Neville, Eds., 2023, pp. 808–816. [Online]. Available: <https://doi.org/10.1609/aaai.v37i1.25159>
- [8] H. Liu, T. Pu, H. Wu, K. Wang, and L. Lin, "DART: dual adaptive refinement transfer for open-vocabulary multi-label recognition," in *ACM International Conference on Multimedia*, 2025, pp. 11 442–11 460.
- [9] L. Chen, W. Zhan, W. Tian, Y. He, and Q. Zou, "Deep integration: A multi-label architecture for road scene recognition," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4883–4898, 2019.
- [10] G. Li, Z. Ji, Y. Chang, S. Li, X. Qu, and D. Cao, "MI-anet: A transfer learning approach using adaptation network for multi-label image classification in autonomous driving," *Chinese Journal of Mechanical Engineering*, vol. 34, 12 2021.
- [11] L. Chen, W. Zhan, W. Tian, Y. He, and Q. Zou, "Deep integration: A multi-label architecture for road scene recognition," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4883–4898, 2019.
- [12] N. Khan, U. Chaudhuri, B. Banerjee, and S. Chaudhuri, "Graph convolutional network for multi-label VHR remote sensing scene recognition," *Neurocomputing*, vol. 357, pp. 36–46, 2019. [Online]. Available: <https://doi.org/10.1016/j.neucom.2019.05.024>
- [13] Q. Song, B. Sun, and S. Li, "DiffCL: A diffusion-based contrastive learning framework with semantic alignment for multimodal recommendations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 10, pp. 17 852–17 862, 2025.
- [14] Q. Song, B. Sun, and S. Li, "Multimodal sparse transformer network for audio-visual speech recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10 028–10 038, 2023.
- [15] J. Tao and X. Fang, "Toward multi-label sentiment analysis: a transfer learning based approach," *J. Big Data*, vol. 7, no. 1, p. 1, 2020. [Online]. Available: <https://doi.org/10.1186/s40537-019-0278-0>
- [16] B. Namrutha Sridhar, K. Mrinalini, and P. Vijayalakshmi, "Data annotation and multi-emotion classification for social media text," in *International Conference on Communication and Signal Processing (ICCCSP)*, 2020, pp. 1011–1015.
- [17] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Int. Conf. Learn. Represent.*, 2018. [Online]. Available: <https://openreview.net/forum?id=rJXMpikCZ>
- [18] Y. Li, D. Tarlow, M. Brockschmidt, and R. S. Zemel, "Gated graph sequence neural networks," in *Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: <http://arxiv.org/abs/1511.05493>
- [19] T. Chen, T. Pu, H. Wu, Y. Xie, L. Liu, and L. Lin, "Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9887–9903, 2022.
- [20] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Int. Conf. Comput. Vis.* IEEE, 2019, pp. 522–531. [Online]. Available: <https://doi.org/10.1109/ICCV.2019.00061>
- [21] K. Du, F. Lyu, L. Li, F. Hu, W. Feng, F. Xu, X. Xi, and H. Cheng, "Multi-label continual learning using augmented graph convolutional network," *IEEE Trans. Multimedia*, vol. 26, pp. 2978–2992, 2024.
- [22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent.*, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [23] H. Wu, W. Chen, Z. Liu, T. Chen, Z. Chen, and L. Lin, "Contrastive transformer learning with proximity data generation for text-based person search," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 7005–7016, 2024.
- [24] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Hcp: A flexible cnn framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, 2015.
- [25] B. Gao and H. Zhou, "Learning to discover multi-class attentional regions for multi-label image recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 5920–5932, 2021. [Online]. Available: <https://doi.org/10.1109/TIP.2021.3088605>
- [26] T. Chen, Z. Wang, G. Li, and L. Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," in *AAAI Conf. Artif. Intell.*, S. A. McIlraith and K. Q. Weinberger, Eds., 2018, pp. 6730–6737. [Online]. Available: <https://doi.org/10.1609/aaai.v32i1.12281>
- [27] X. Li, S. Liao, W. Lan, X. Du, and G. Yang, "Zero-shot image tagging by hierarchical semantic embedding," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, R. Baeza-Yates, M. Lalmas, A. Moffat, and B. A. Ribeiro-Neto, Eds., 2015, pp. 879–882. [Online]. Available: <https://doi.org/10.1145/2766462.2767773>
- [28] C. Lee, W. Fang, C. Yeh, and Y. F. Wang, "Multi-label zero-shot learning with structured knowledge graphs," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1576–1585.
- [29] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, A. Moschitti, B. Pang, and W. Daelemans, Eds., 2014, pp. 1532–1543.
- [30] D. Huynh and E. Elhamifar, "A shared multi-attention framework for multi-label zero-shot learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 8773–8783.
- [31] S. Narayan, A. Gupta, S. H. Khan, F. S. Khan, L. Shao, and M. Shah, "Discriminative region-based multi-label zero-shot learning," in *Int. Conf. Comput. Vis.*, 2021, pp. 8711–8720.
- [32] L. H. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint*, vol. abs/1908.03557, 2019. [Online]. Available: <http://arxiv.org/abs/1908.03557>
- [33] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Adv. Neural Inform. Process. Syst.*, vol. 34, pp. 9694–9705, 2021.
- [34] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei, "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," *Adv. Neural Inform. Process. Syst.*, vol. 35, pp. 32 897–32 912, 2022.
- [35] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 14 393–14 402.
- [36] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Open-vocabulary detr with conditional matching," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 106–122.

- [37] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 14 084–14 093.
- [38] D. Huynh, J. Kuen, Z. Lin, J. Gu, and E. Elhamifar, "Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 7020–7031.
- [39] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 540–557.
- [40] H. Liu, Z. Song, H. Wu, T. Pu, K. Wang, and L. Lin, "Robust egocentric referring video object segmentation via dual-modal causal intervention," in *Annual Conference on Neural Information Processing Systems*, 2025.
- [41] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "PLA: language-driven open-vocabulary 3d scene understanding," in *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, 2023, pp. 7010–7019. [Online]. Available: <https://doi.org/10.1109/CVPR52729.2023.00677>
- [42] S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. A. Funkhouser, "Openscene: 3d scene understanding with open vocabularies," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17–24, 2023.* IEEE, 2023, pp. 815–824. [Online]. Available: <https://doi.org/10.1109/CVPR52729.2023.00085>
- [43] H. Wu, Y. Chen, L. Liu, T. Chen, K. Wang, and L. Lin, "SQLNet: Scale-modulated query and localization network for few-shot class-agnostic counting," *IEEE Transactions on Image Processing*, vol. 34, pp. 4631–4645, 2025.
- [44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.
- [45] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Int. Conf. Mach. Learn.* PMLR, 2021, pp. 4904–4916.
- [46] X. Gu, T. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *Int. Conf. Learn. Represent.*, 2022. [Online]. Available: <https://openreview.net/forum?id=IL3lnMbR4WU>
- [47] Z. Ma, G. Luo, J. Gao, L. Li, Y. Chen, S. Wang, C. Zhang, and W. Hu, "Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 14 074–14 083.
- [48] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: <https://doi.org/10.18653/v1/n19-1423>
- [49] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [50] Z. Huang, G. Ling, S. Zhong, H. Wu, and L. Lin, "Minilongbench: The low-cost long context understanding benchmark for large language models," in *Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 11 442–11 460.
- [51] Z. Huang, G. Ling, V. S. Liang, Y. Lin, Y. Chen, S. Zhong, H. Wu, and L. Lin, "Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms," in *Conference on Empirical Methods in Natural Language Processing, Findings*, 2025.
- [52] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" in *Int. Conf. Learn. Represent.*, 2022. [Online]. Available: <https://openreview.net/forum?id=F72ximsx7C1>
- [53] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, pp. 3538–3545.
- [54] K. Oono and T. Suzuki, "Graph neural networks exponentially lose expressive power for node classification," in *International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: <https://arxiv.org/abs/1905.10947>
- [55] G. Li, M. Muller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs go as deep as CNNs?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [56] S. Rahman and S. Khan, "Deep multiple instance learning for zero-shot image tagging," in *Asian Conf. Comput. Vis.* Springer, 2018, pp. 530–546.
- [57] Z. Liu, S. Guo, X. Lu, J. Guo, J. Zhang, Y. Zeng, and F. Huo, "(ml)²-encoder: On exploration of channel-class correlation for multi-label zero-shot learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 23 859–23 868.
- [58] A. Ben-Cohen, N. Zamir, E. Ben-Baruch, I. Friedman, and L. Zelnik-Manor, "Semantic diversity learning for zero-shot multi-label classification," in *Int. Conf. Comput. Vis.*, 2021, pp. 640–650.
- [59] Y. Lin, M. Chen, K. Zhang, H. Li, M. Li, Z. Yang, D. Lv, B. Lin, H. Liu, and D. 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*, vol. 38, no. 4, 2024, pp. 3513–3521.
- [60] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: a real-world web image database from national university of singapore," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, S. Marchand-Maillet and Y. Kompatsiaris, Eds., 2009.
- [61] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1956–1981, 2020.
- [62] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [63] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *arXiv preprint arXiv:2310.03744*, 2023.
- [64] LLaMA 3 Teams, "The Llama 3 herd of models," 2024. [Online]. Available: <https://arxiv.org/abs/2407.21783>
- [65] Qwen Team, "Qwen2.5: A party of foundation models," Sept. 2024. [Online]. Available: <https://qwenlm.github.io/blog/qwen2.5/>
- [66] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Int. Conf. Comput. Vis.* IEEE, 2017, pp. 618–626. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.74>