# Language-Driven Cross-Modal Classifier for Zero-Shot Multi-Label Image Recognition

**Yicheng Liu** [1] **Jie Wen** [*1] **Chengliang Liu** [1] **Xiaozhao Fang** [*2] **Zuoyong Li** [3] **Yong Xu** [1] **Zheng Zhang** [*1]

## Abstract

Large-scale pre-trained vision-language models (e.g., CLIP) have shown powerful zero-shot transfer capabilities in image recognition tasks. Recent approaches typically employ supervised fine-tuning methods to adapt CLIP for zero-shot multi-label image recognition tasks. However, obtaining sufficient multi-label annotated image data for training is challenging and not scalable. In this paper, we propose a new language-driven framework for zero-shot multi-label recognition that eliminates the need for annotated images during training. Leveraging the aligned CLIP multi-modal embedding space, our method utilizes language data generated by LLMs to train a cross-modal classifier, which is subsequently transferred to the visual modality. During inference, directly applying the classifier to visual inputs may limit performance due to the modality gap. To address this issue, we introduce a cross-modal mapping method that maps image embeddings to the language modality while retaining crucial visual information. Comprehensive experiments demonstrate that our method outperforms other zero-shot multi-label recognition methods and achieves competitive results compared to few-shot methods.

## 1. Introduction

Large-scale multi-modal pre-trained models, such as Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021), have shown impressive capabilities in cross-modal representation learning. CLIP leveraged contrastive learning from 400 million image-text pairs to map images and texts into a shared embedding space. Such models show remarkable generalization capabilities in various downstream vision tasks (Zhou et al., 2022; Zhao et al., 2022; Luo et al., 2023; Wang et al., 2023; Zhang et al., 2024).

Multi-label recognition (MLR) is an important vision task, which aims to describe what is present in an image using multiple labels. Compared to single-label image recognition tasks, which solely focus on the subject objects in an image, MLR tasks must recognize images with more complex scenarios and multiple objects. Many works have achieved great improvements in adopting CLIP to multi-label recognition tasks. For instance, ADDS (Xu et al., 2022) learned a transformer decoder to facilitate the fusion of the semantics from dual-modal information sources. DualCoOp (Sun et al., 2022) learned positive and negative prompts to adapt the knowledge learned in CLIP to multi-label image recognition. HSPNet (Ramesh et al., 2022) proposed a hierarchical semantic prompt network to explore the hierarchical semantic relationship in the CLIP model. Although these methods can achieve remarkable performance of multi-label image recognition, they usually require sufficient labeled images to fine-tune the model. Unfortunately, the collection of large-scale and high-quality annotated multi-label datasets remains challenging and resource-intensive.

To mitigate this issue, researchers have developed a new setting, *i.e.*, multi-label zero-shot learning (ML-ZSL), in which an image is associated with potentially multiple seen and unseen classes, but only labels of seen classes are provided during training. Recent methods (Zhang et al., 2016; Ben-Cohen et al., 2021; He et al., 2023; Pu et al., 2023; Chen et al., 2022) commonly utilized label embeddings in the semantic space to transfer the knowledge from seen classes to unseen classes. However, these methods still require image data annotated by seen classes to train the model and suffer from significant performance degradation due to the lack of annotated image data.

Generally speaking, compared with the collection of large-scale annotated image datasets, it is significantly easier to collect large amounts of natural language data. If we can find a cohesive alignment of semantics across vision and language modalities, then we can explore the information

[*]Corresponding author [1]Harbin Institute of Technology, Shenzhen, China [2]Guangdong University of Technology, Guangzhou, China [3]Minjiang University, Fuzhou, China. Correspondence to: Jie Wen <jiewen_pr@126.com>, Xiaozhao Fang <xzhfang168@126.com>, Zheng Zhang <darrenzz219@gmail.com>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

of language data to promote the performance of MLR in the vision space. Inspired by this motivation, to diminish the reliance on collecting large-scale annotated image datasets, we propose a new zero-shot multi-label image recognition framework using language-only data for training. The proposed framework is illustrated in Figure 1. Specifically, to collect sufficient high-quality language data, we exploit a Large Language Model (LLM) such as GPT-3 (Brown et al., 2020) to generate textual descriptions for all image categories in a dataset. Compared with collecting textual data on the web, generating data using LLM is more straightforward and requires less manual post-processing. Considering that only training the model with language data will lead to a modality gap phenomenon between texts and images, we further propose a simple and effective mapping method to transfer the image embedding into the text embedding space to reduce the impact of the modality gap. In particular, we set up a series of label-specific text embeddings and represent the image embedding as a linear combination of text embeddings. At the same time, we retain the important visual information by fusing the mapped embedding and the original image embedding. Additionally, we take full advantage of the powerful CLIP model to extract both global and local image information, enhancing our classifier's ability to recognize multiple objects in an image. The main contributions can be summarized as follows:

- We propose a new language data-driven framework for zero-shot multi-label recognition. The method can train a **Cr**oss-**M**odal **C**lassifier (**CoMC**) with language data for the multi-label image recognition tasks, which efficiently eliminates the requirement of large-scale high-quality annotated image data.
- We propose a simple yet efficient cross-modal mapping method to align the text and vision embeddings, which can effectively reduce the modality gap.
- Extensive experiment results show that without using image data to train the model, our method still performs significantly better than many zero-shot methods and few-shot methods for MLR.

## 2. Related Work

**Multi-label Zero-shot Learning.** Zero-shot learning aims to train a model for classifying objects of unseen classes. Most studies focus on single-label recognition tasks, which classify an image into one category (Novack et al., 2023; Pratt et al., 2023; Naeem et al., 2023). Although these methods have achieved significant success, they cannot be well transferred to zero-shot multi-label recognition tasks, which is more challenging and more practical in real-world applications. In most early works, labels were split into seen and unseen categories, and models trained on images with seen labels were applied to unseen labels during inference. LabelEM (Akata et al., 2015) introduced a function that measures the compatibility between image and label to learn a joint image-label embedding. Fast0Tag (Zhang et al., 2016) and SDL (Ben-Cohen et al., 2021) learned one or multiple diverse principal embedding vectors of the image. Generally speaking, multiple objects are usually distributed across multiple regions of an image. To obtain the relevant object regions, LESA (Huynh & Elhamifar, 2020) designed a shared multi-attention mechanism. Deep0Tag (Rahman et al., 2019) learned a region proposal network to automatically locate relevant image patches. BiAM (Narayan et al., 2021) enhanced the region-based features by minimizing the inter-class feature entanglement. Though significant progress has been made, existing methods still overly rely on large-scale annotated image data and generally have complex network architectures or loss functions.

**Vision-language Pre-trained Models.** With the increase of large-scale visual text pairs collected from the Internet, visual language pretraining (VLP) has become a hot research topic in recent years (Luo et al., 2020; Huang et al., 2020; Kim et al., 2021; Li et al., 2022; Jia et al., 2021; Radford et al., 2021). Among them, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) learned visual and textual representations from millions of image-text pairs collected from the Internet, showing superior zero-shot transferability in various downstream tasks. To further improve the VLP performance, some works (Mu et al., 2022; Li et al., 2021) utilized additional self-supervision within-modality. FILIP (Yao et al., 2021) considered capturing finer-grained image-text relationships. For the above methods, fine-tuning is necessary when adopting these VL models to downstream tasks. However, directly fine-tuning the whole model consumes a lot of computational resources and may compromise the original representation learning capability of the model. To address this issue, CoOp (Zhou et al., 2022) introduced the concept of prompt tuning to the vision domain, providing an efficient way to fine-tune the VL models. Furthermore, DualCoOp (Sun et al., 2022) learned dual prompts to adapt CLIP to multi-label learning tasks. For the above methods, a common limitation is that these methods require annotated image data to fine-tune the VL models to ensure their performance. TaI-DPT (Guo et al., 2023) treated text data as images for prompt tuning. While both our method and TaI-DPT use text data for training, we distinguish ourselves by leveraging the aligned embedding space of VL models to train a cross-modal classifier using text embeddings directly.

## 3. Method

The proposed framework is shown in Figure 1. First, we adopt an LLM to generate a diverse multi-label text dataset. Then, we use the CLIP text encoder to encode the input

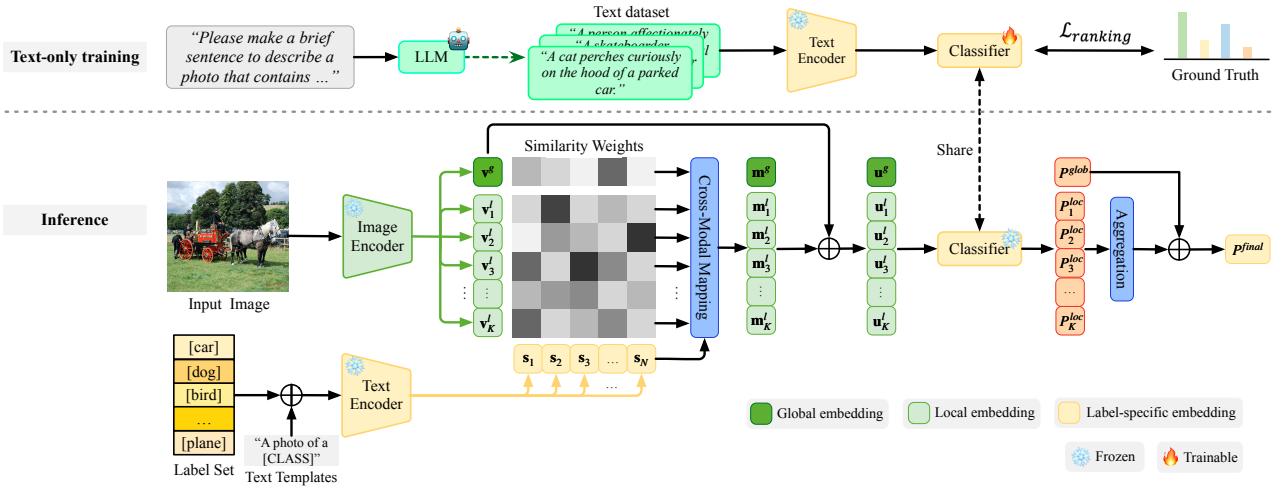



*Figure 1.* An overview of our framework. Our method is based on a pre-trained contrastive model CLIP containing a visual encoder and a text encoder. (**Top**) During training, we learned a classifier for multi-label recognition based on the CLIP text embedding. The pre-trained CLIP image and text encoders are kept frozen. (**Bottom**) At inference, we use a simple mapping method to map the image embedding into text embedding space with the help of a series of label-specific text embeddings. Then the mapped embedding is fused with the original image embedding for multi-label recognition.

text, and then train a multi-label classifier to recognize the textual data. Thanks to the aligned image-text embedding space of CLIP, we can transfer the classifier trained by texts to zero-shot image recognition. At inference, to enhance the discriminative ability to recognize multiple objects appearing in different regions of the image, global and local visual embeddings are both extracted using the CLIP image encoder. Furthermore, we propose a simple yet efficient mapping method to reduce the modality gap between the text embedding space and the image embedding space. Then the mapped embedding is fed into the multi-label classifier for image recognition.

### 3.1. LLM-based Multi-label Text Data Generation

As mentioned in the previous section, we attempt to train a language-driven network to address the dependence issue on large-scale annotated images. Therefore, the first work is to collect the high-quality language data associated with the multi-label training data. Directly collecting and labeling text data from the web can be labor-intensive, and the collected data may prove unusable if it lacks sufficient diversity. Large Language Models (LLM), such as GPT-3 (Brown et al., 2020), offer a more efficient approach to acquiring high-quality textual training data.

We leverage GPT-3 to generate the text data for the training of our multi-label classifier. Specifically, we first sample several labels randomly from the label set of the target dataset. To guide GPT-3 in generating concise textual descriptions, we employ the following LLM-prompt: '*Please make a brief sentence to describe a photo that contains ...*', filled in with the label names we sampled. This process is iterated to generate 40,000 sentences covering various combinations of labels. This approach allows us to acquire a large-scale high-quality text dataset efficiently. Leveraging this generated text dataset, we proceed to train our multi-label classifier.

### 3.2. Text-only Training of Cross-modal Classifier

The pre-trained vision-language models generally adopt contrastive learning to find an embedding alignment space for image and text modalities. As proved in (Zhang et al., 2022), such shared embedding space learned by these pre-trained multi-modal models makes the cross-modal transfer possible. In other words, it is possible to replace the image inputs with text inputs as good proxies to promote the performance of pre-trained vision-based models. Inspired by this motivation, we try to train a multi-label image classifier using text data.

In our work, we simply use a single linear classifier to perform multi-label recognition. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ be the label set of the target dataset, where $N$ is the number of labels. The text training set is denoted as $\mathcal{D} = \{(\mathcal{T}_i, \mathcal{Y}_i)\}_{i=1}^{M}$, where $M$ is the number of texts; $\mathcal{Y}_i = \{y_{i,1}, y_{i,2}, \ldots, y_{i,N}\}$ denotes the ground truth labels of the text $\mathcal{T}_i$; $y_{i,j}$ for $j \in \{1, 2, \ldots, N\}$ is 1 if the text $\mathcal{T}_i$ is generated from the label $c_j$ and 0 otherwise. Given a text $\mathcal{T}_i$, we use the CLIP text encoder to extract its embedding, formulated as follows:

$$\mathbf{t}_i = E_t(\mathcal{T}_i) \tag{1}$$

where $\mathbf{t}_i \in \mathbb{R}^d$ is the text embedding, $d$ is the dimension of text embedding; $E_t(\cdot)$ denotes the CLIP text encoder. Then the $l_2$ normalized text embedding is fed into the multi-label classifier $f_\theta$ to compute the confidence score:

$$P_i = f_\theta(\frac{\mathbf{t}_i}{\|\mathbf{t}_i\|_2}) \tag{2}$$

where $P_i = \{p_{i,1}, p_{i,2}, \ldots, p_{i,N}\}$ and $p_{i,j}$ for $j \in \{1, 2, \ldots, N\}$ being the probability that the text $\mathcal{T}_i$ is generated from label $c_j$. Following (Guo et al., 2023), we use ranking loss $\mathcal{L}_{ranking}$ to measure the discrepancy between confidence scores and ground-truth labels. Formally, the training objective is:

$$\mathcal{L}_{ranking} = \sum_{j \in \{c^+\}} \sum_{k \in \{c^-\}} \max(0, m - (p_{i,j} - p_{i,k})) \tag{3}$$

where $\{c^+\}$ and $\{c^-\}$ denote the positive classes and negative classes, respectively. $m$ is the margin value to determine how much larger the similarity score between positive classes should be than between negative classes. We set the margin value $m = 1$ in ranking loss.

The detailed processes to train a language-driven cross-modal classifier are summarized in Algorithm 1.

**Algorithm 1** Training process of CoMC.

**Input:** Text dataset $\mathcal{D} = \{(\mathcal{T}_i, \mathcal{Y}_i)\}_{i=1}^M$, classifier $f_\theta$, text encoder $E_t(\cdot)$, learning rate $\delta$, training epochs $T$, the number of iteration $I$, batch size $B$.
**Initialize:** Text encoder $E_t(\cdot)$ is initialized as the parameters in the pre-trained CLIP. The parameter $\theta$ of the classifier is randomly initialized.
**for** $t = 1$ **to** $T$ **do**
  **for** $k = 1$ **to** $I$ **do**
    Sample batch $\{(\mathcal{T}_i, \mathcal{Y}_i)\}_{i=1}^B \subset \mathcal{D}$;
    **for** $i = 1$ **to** $B$ **do**
      Extract the text embedding $\mathbf{t}_i$ of text $\mathcal{T}_i$ according to Eq.(1);
      Calculate the confidence scores $P_i$ using Eq.(2);
    **end for**
    Calculate the ranking loss $\mathcal{L}_{ranking}$ using Eq.(3) and update the parameters of classifier $\theta$ using Adam optimizer.
  **end for**
**end for**
**Output:** The parameters of the classifier $\theta$.

### 3.3. Inference Stage

After training a multi-label classifier using textual descriptions generated by LLM, now we have to apply this classifier to image embeddings to achieve cross-modal transfer.

**Fine-grained Image Embeddings**. Vision-language models based on contrastive learning have demonstrated impressive success in zero-shot vision recognition tasks. However, CLIP (Radford et al., 2021) only focuses on matching each image with a single label during its training, hence it is not suitable to handle the multi-label recognition cases. In addition, standard CLIP only uses a global image embedding to align the image to a single label. As a result, its embedding can only describe the most dominant objects in the image while ignoring other objects in the image. To better distinguish multiple objects in the MLR task, we propose to extract both global embedding $\mathbf{v}^g \in \mathbb{R}^d$ and the flattened feature map $\mathbf{F} \in \mathbb{R}^{K \times d}$ before the attention pooling layer of CLIP image encoder. $d$ is the dimension of CLIP embedding. $K = H \times W$ denotes the spatial dimension of the flattened feature map, where $H$ and $W$ are the height and width of the feature map, respectively. We then split $\mathbf{F}$ into multiple local embeddings along the spatial dimension, and get local embeddings set $\mathbf{V}^l = \{\mathbf{v}_1^l, \mathbf{v}_2^l, \ldots, \mathbf{v}_K^l\}$, where $\mathbf{v}_j^l \in \mathbb{R}^d$ denotes the local embedding of the spatial region $j$.

The two kinds of feature embeddings describe the image more comprehensively, which enables our method to obtain a more discriminative classifier yet better performance on multi-label/multi-object recognition tasks.

**Mapping Image-to-Text**. Our target is to train a classifier with language-only data that can be transferred to the vision modality. Therefore, the encoded text embeddings should match the corresponding images. However, as shown in (Liang et al., 2022), although the pre-trained contrastive model CLIP aligned the text embedding and image embedding, there still exists a modality gap between the text embedding space and image embedding space. It may limit the performance of the multi-label image recognition if we directly feed the image embedding into the classifier trained by the text embedding. (Ramesh et al., 2022) establish a relationship between images and texts in a supervised manner. However, it requires a large amount of paired image-text data to train the model. To reduce the influence of the modality gap, we propose a simple mapping method to map the image embedding into the text embedding space.

Specifically, different from the conventional mapping methods, such as adding contrastive loss or mapping networks, we try to represent the image embeddings as the linear combination of a series of text embeddings, where the linear combination weights are computed according to the similarity relationships of the image embedding and label-specific text embedding. First, following (Radford et al., 2021), we construct a text template '`a photo of a/an [c].`' for each class '`c`' in the target image dataset. Then the text templates are fed into the CLIP text encoder to extract feature set $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$, where $N$ denotes

the number of classes. $\mathbf{s}_i$ can be viewed as a label-specific text embedding for the $i$-th class $c_i$. We compute the cosine similarity between image embedding (including the global embedding and local embedding) and text embedding corresponding to each class in $\mathbf{S}$ and then normalize it using softmax to obtain the weights for each label-specific text embedding. Then the mapped global and local embeddings in the text embedding space are calculated as:

$$\mathbf{m}^g = \sum_{i=1}^{N} w_i^g \cdot \mathbf{s}_i = \sum_{i=1}^{N} \frac{\exp(\mathbf{s}_i^\top \mathbf{v}^g/\tau)}{\sum_{i=1}^{N} \exp(\mathbf{s}_i^\top \mathbf{v}^g/\tau)} \cdot \mathbf{s}_i \quad (4)$$

$$\mathbf{m}_j^l = \sum_{i=1}^{N} w_{i,j}^l \cdot \mathbf{s}_i = \sum_{i=1}^{N} \frac{\exp(\mathbf{s}_i^\top \mathbf{v}_j^l/\tau)}{\sum_{i=1}^{N} \exp(\mathbf{s}_i^\top \mathbf{v}_j^l/\tau)} \cdot \mathbf{s}_i \quad (5)$$

where $\mathbf{m}^g$ and $\mathbf{m}_j^l$ denote the mapped global image embedding and local image embedding, respectively. $\tau$ is the temperature parameter. And the mapped local image embedding set $M^l = \{\mathbf{m}_j^l\}_{j=1}^K$. These label-specific text embeddings are computed before the inference stage and cached in memory.

For the language-driven models, another issue is that the mapped embedding may lose some important visual information owing to the simple selected text templates. To address this issue, we further introduce a weighted fusion strategy to retain the original visual information as follows:

$$\mathbf{u}^g = \alpha \mathbf{v}^g + (1-\alpha)\mathbf{m}^g \quad (6)$$

$$\mathbf{u}_j^l = \alpha \mathbf{v}_j^l + (1-\alpha)\mathbf{m}_j^l \quad (7)$$

where $\alpha$ is a weighting factor. $\mathbf{u}^g$ and $\mathbf{u}_j^l$ represent the fused global image embedding and fused local image embedding, respectively. The two fused embeddings are then fed into the multi-label classifier $f_\theta$ to compute the confidence scores of all labels as follows:

$$P^{glob} = f_\theta(\mathbf{u}^g) \quad (8)$$

$$P_j^{loc} = f_\theta(\mathbf{u}_j^l) \quad (9)$$

where $P^{glob} = \{p_1^{glob}, p_2^{glob}, \ldots, p_N^{glob}\}$ is the global confidence score and $P_j^{loc} = \{p_{j,1}^{loc}, p_{j,2}^{loc}, \ldots, p_{j,N}^{loc}\}$ denotes the local confidence scores for each spatial region $j$. Now each input image is associated with one global confidence score $P^{glob}$ and $K$ local confidence scores $P_j^{loc}$. To aggregate these local confidence scores, we extract the highest confidence score that class $i$ obtains among all local regions. Let

$$q_i = \max_{j=1,\ldots,K} p_{j,i}^{loc} \quad (10)$$

The aggregation score for all classes is given by:

$$P^{agg} = \{q_1, q_2, \ldots, q_N\} \quad (11)$$

*Table 1.* Comparison with zero-shot learning methods without image training on MS-COCO, VOC2007, and NUS-WIDE. The evaluation is based on mAP (%).

| Method | MS-COCO | VOC2007 | NUS-WIDE |
|---|---|---|---|
| Zero-shot CLIP | 47.3 | 76.2 | 36.4 |
| CLIP-DPT | 49.7 | 77.3 | 37.4 |
| TaI-DPT | 65.1 | 88.3 | 46.5 |
| CoMC | **68.7** | **89.4** | **48.2** |

Finally, $P^{glob}$ and $P^{agg}$ are combined by calculating the average to obtain the final predicted probability for all labels as follows:

$$P^{final} = \frac{1}{2}(P^{glob} + P^{agg}) \quad (12)$$

The detailed processes of the inference stage are summarized in Algorithm 2. Our cross-modal mapping method can efficiently map the image embedding to the text embedding space, reducing the impact of the modality gap. Meanwhile, it retains the visual information of the original image embedding for visual recognition. When adapting to different datasets, we can just simply replace the classes in the text template with the classes of the target dataset without additional training.

**Algorithm 2** Inference process of CoMC.

**Input:** Test image $\mathcal{I}$, Classifier $f_\theta$, Image encoder $E_v(\cdot)$, Updated parameters $\theta$ of the classifier.
**Initialize:** Image encoder $E_v(\cdot)$ is initialized by the pre-trained CLIP parameters. Classifier $f_\theta$ is initialized by $\theta$.
Extract the global embedding $\mathbf{v}^g$ and local embedding set $\{\mathbf{v}_i^l\}_{i=1}^K$ by $E_v(\cdot)$;
Calculate the mapped embeddings $\mathbf{m}^g$ and $\{\mathbf{m}_i^l\}_{i=1}^K$ using Eq.(4)(5);
Calculate the fuse embeddings $\mathbf{u}^g$ and $\{\mathbf{u}_i^l\}_{i=1}^K$ using Eq.(6)(7);
Compute the global confidence score $P^{glob}$ and local confidence scores $\{P_i^{loc}\}_{i=1}^K$ using Eq.(8)(9);
Aggreate the local confidence scores to $P^{agg}$ using Eq.(10);
$P^{final} = \frac{1}{2}(P^{glob} + P^{agg})$.
**Output:** The final predicted probability $P^{final}$.

## 4. Experiment

### 4.1. Experimental Setup

**Datasets.** We conduct experiments on MS-COCO (Lin et al., 2014), VOC2007 (Everingham et al., 2010), and

*Table 2.* Comparison with related multi-label zero-shot learning methods with image training on the NUS-WIDE dataset. We report the results in terms of mAP, as well as precision (**P**), recall (**R**), and **F1** score at $K \in \{3, 5\}$.

| Method | Top-3 | | | Top-5 | | | mAP |
|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** | |
| CONSE (Norouzi et al., 2013) | 17.5 | 28.0 | 21.6 | 13.9 | 37.0 | 20.2 | 9.4 |
| LabelEM (Akata et al., 2015) | 15.6 | 25.0 | 19.2 | 13.4 | 35.7 | 19.5 | 7.1 |
| Fast0Tag (Zhang et al., 2016) | 22.6 | 36.2 | 27.8 | 18.2 | 48.4 | 26.4 | 15.1 |
| One Attention per Label (Kim et al., 2018) | 20.9 | 33.5 | 25.8 | 16.2 | 43.2 | 23.6 | 10.4 |
| LESA (M=10) (Huynh & Elhamifar, 2020) | 25.7 | 41.1 | 31.6 | 19.7 | 52.5 | 28.7 | 19.4 |
| BiAM (Narayan et al., 2021) | - | - | 33.1 | - | - | 30.7 | 26.3 |
| SDL (M=7) (Ben-Cohen et al., 2021) | 24.2 | 41.3 | 30.5 | 18.8 | 53.4 | 27.8 | 25.9 |
| MKT (He et al., 2023) | 27.7 | 44.3 | 34.1 | 21.4 | 57.0 | 31.1 | 37.6 |
| DualCoOp (Sun et al., 2022) | **37.3** | 46.2 | **41.3** | **28.7** | 59.3 | **38.7** | 43.6 |
| CoMC | 33.5 | **53.5** | 41.2 | 24.8 | **66.1** | 36.1 | **48.2** |

*Table 3.* Comparison with multi-label few-shot methods on VOC2007 and MS-COCO. The evaluation is based on mAP (%) for 0-shot, 1-shot, 2-shot, 4-shot, 8-shot, and 16-shot with treating all classes as novel classes.

| Method | VOC2007 | | | | | | MS-COCO | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0-shot | 1-shot | 2-shot | 4-shot | 8-shot | 16-shot | 0-shot | 1-shot | 2-shot | 4-shot | 8-shot | 16-shot |
| CoOp | - | 79.3 | 83.2 | 83.8 | 84.5 | 85.7 | - | 52.6 | 57.3 | 58.1 | 59.2 | 59.8 |
| CoOp-DPT | - | 83.2 | 88.1 | 88.2 | 90.0 | 90.1 | - | 65.8 | 66.2 | 67.6 | 68.1 | 68.9 |
| CoMC | **89.4** | **89.7** | **90.1** | **90.6** | **91.4** | **92.1** | **68.7** | **68.9** | **69.3** | **70.4** | **70.9** | **71.4** |

NUS-WIDE (Chua et al., 2009) to evaluate the superiority of the proposed method on multi-label recognition tasks. As our method does not use any image for training, we adopt their official test set to evaluate our method. Specifically, MS-COCO contains 80 categories, and we take the official val2014 (40K images) splits for testing. VOC2007 contains 20 categories and we use the official test (5K images) splits for testing. NUS-WIDE contains 161,789 training images from 81 categories and we use the remaining 107,859 images as test samples.

**Implementation Details.** For a fair comparison, we use CLIP ResNet-50 as the image encoder and CLIP Transformer as the text encoder. The multi-label classifier mainly contains a single linear layer, where the input units are equal to the embedding dimension of the CLIP encoder and the output units are equal to the number of classes in the target dataset. During training, we keep the two CLIP encoders frozen and only train the parameters of the linear classifier. We use a cosine learning rate decay with an initial learning rate of 1e-4. We train our classifier using the Adam optimizer with a batch size of 256 to optimize the classifier for 30 epochs. For inference, the input images are resized into $224 \times 224$. The text templates are fed into the CLIP text encoder to produce their feature embeddings. The weighting factor $\alpha$ is set to 0.7, 0.4, and 0.5 on MS-COCO, VOC2007, and NUS-SIDE, respectively. The temperature parameter $\tau$ is set to 1/100. For text data generation, GPT-3 DaVinci-002 (Brown et al., 2020) model is adopted, where 40000 textual descriptions are generated for each dataset.

### 4.2. Experimental Results and Analysis

#### 4.2.1. Comparsion with Zero-shot Learning Methods

In the traditional zero-shot multi-label recognition task, datasets are usually split into seen classes and unseen classes. Previous works use the seen classes for training and then predict the unseen classes during inference. In our work, there is no labeled image data used for training, so we compared our work with methods not using image training and methods using image training, respectively.

**Comparison with Methods without Image Training**. We compare our method with the following baselines: Zero-shot CLIP (Radford et al., 2021), CLIP-DPT (Guo et al., 2023), and TaI-DPT (Guo et al., 2023). Table 1 shows the zero-shot multi-label recognition performance of the above three methods and our method on the three datasets. We can observe that under the zero-shot setting, the performance of our method on the three datasets is 3.6%, 1.1%, and 1.7% higher than the top-1 ranked method TaI-DPT (Guo et al., 2023), respectively, showing the effectiveness of our method. Different from TaI-DPT which uses caption data from public image caption datasets, our method adopts the

*Table 4.* Ablation study on the main component of CoMC. Local-emb refers to local image embeddings, Mapping refers to cross-modal mapping. The mAP (%) values on MS-COCO, VOC2007, and NUS-WIDE are reported.

| Model | Local-emb | Mapping | MS-COCO | VOC2007 | NUS-WIDE |
|---|---|---|---|---|---|
| CoMC | ✗ | ✗ | 65.9 | 86.4 | 45.0 |
| CoMC | ✗ | ✓ | 66.2 | 88.5 | 47.6 |
| CoMC | ✓ | ✗ | 67.9 | 87.5 | 45.1 |
| CoMC | ✓ | ✓ | **68.7** | **89.4** | **48.2** |

*Table 5.* Effect of different LLMs for generating datasets. The evaluation is based on mAP (%).

| Method | MS-COCO | VOC2007 | NUS-WIDE |
|---|---|---|---|
| Zero-shot CLIP | 47.3 | 76.2 | 36.4 |
| CoMC (ChatGLM-6b) | 66.2 | 85.1 | 42.8 |
| CoMC (Llama-2-7b) | 66.1 | 87.5 | 46.9 |
| CoMC (GPT-3) | **68.7** | **89.4** | **48.2** |

text data generated by LLM for training and assisting in image recognition. This illustrates that the text data generated by LLM can cover a richer combination of categories. Moreover, compared with the original Zero-shot CLIP, our method surpasses it by a large margin. This also shows that the original CLIP classifier of computing similarity between images and texts does not work well on the MLR task.

**Comparison with Methods with Image Training**. We compare our method with the following eight baselines: CONSE (Norouzi et al., 2013), LabelEM (Akata et al., 2015), Fast0Tag (Zhang et al., 2016), One Attention per Label (Kim et al., 2018), LESA (Huynh & Elhamifar, 2020), BiAM (Narayan et al., 2021), SDL (Ben-Cohen et al., 2021), MKT (He et al., 2023), and DualCoOp (Sun et al., 2022). We follow (Sun et al., 2022) to report mAP over all categories as well as precision, recall, and F1 score at Top-3 and Top-5 predictions in each image on the NUS-WIDE dataset. Table 2 shows the experimental results of our method and the above nine methods on the NUS-WIDE datasets at the zero-shot learning case. From the table, we can observe that our model achieves the highest mAP value which is 4.6% higher than the second-best method DualCoOp. MKT is also a CLIP-based method, our method surpasses MKT with an absolute gain of 10.6% mAP and improves the F1 score by absolute gains of 7.1% and 5.0% at $K = 3$ and $K = 5$, respectively. Moreover, our method does not use any images for model training, demonstrating our its superiority.

#### 4.2.2. Comparsion with Few-shot Learning Methods.

Following experimental settings in (Guo et al., 2023), we treat all classes as novel classes and select 1, 2, 4, 8, and 16 shot samples for each class for training. To implement our CoMC in a few-shot setting, we use the labeled samples to fine-tune our classifier after training with text data. In Table 3, we provide a comparison of our CoMC and two few-shot learning methods CoOp and CoOp-DPT, where CoOp-DPT is an extension of CoOp proposed by (Guo et al., 2023). We see that zero-shot CoMC is 1.2% better than 4-shot CoOp-DPT on VOC2007 and 0.6% better than 8-shot CoOp-DPT on MS-COCO. Besides, the performance of CoMC gets a stable improvement on two datasets as the number of labeled samples increases, this also demonstrates the effectiveness of our CoMC.

*Table 6.* Effect of different CLIP visual backbones. The mAP (%) values of CLIP, CLIP-DPT, and the proposed CoMC are reported.

| Backbone | Method | MS-COCO | VOC2007 | NUS-WIDE |
|---|---|---|---|---|
| RN50 | CLIP | 47.3 | 76.2 | 36.4 |
| | CLIP-DPT | 49.7 | 77.3 | 37.4 |
| | CoMC | **68.7** | **89.4** | **48.2** |
| RN101 | CLIP | 48.6 | 76.8 | 37.2 |
| | CLIP-DPT | 54.1 | 81.3 | 38.3 |
| | CoMC | **71.2** | **90.0** | **48.8** |
| ViT-B/32 | CLIP | 49.4 | 76.6 | 37.8 |
| | CLIP-DPT | 55.2 | 82.7 | 38.9 |
| | CoMC | **71.9** | **89.9** | **48.6** |

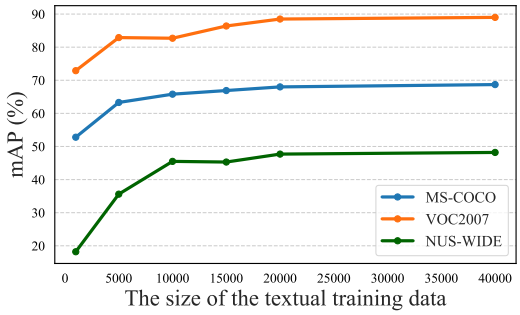



*Figure 2.* Effect of the size of textual training data on MS-COCO, VOC2007, and NUS-WIDE. The mAP (%) is reported.

### 4.3. Ablation Study

**Ablation of Cross-modal Mapping.** To reduce the impact of the modality gap, we design a simple cross-modal mapping method to map the image embedding into text embedding space. To verify the effectiveness of cross-modal mapping, we wipe out cross-modal mapping in our method. The result is shown in Table 4, without cross-modal mapping, the performance of the model on three datasets decreases substantially. This indicates the importance of bridging the modality gap between images and texts and demonstrates the effectiveness of our designed linear representation and fusion-based cross-modal mapping approach.

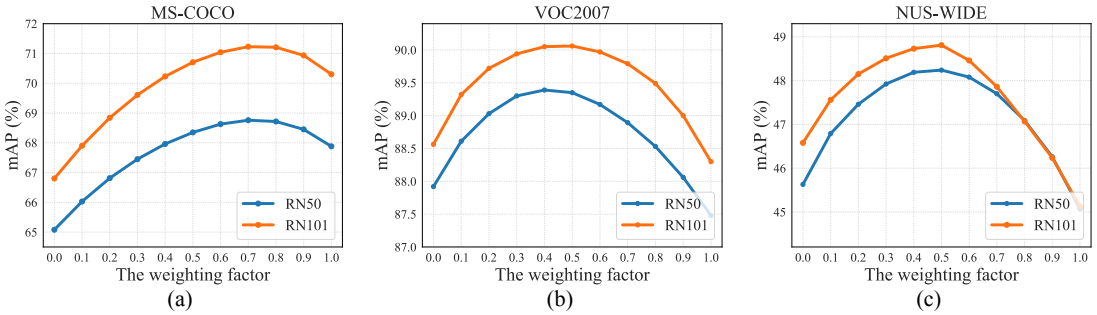
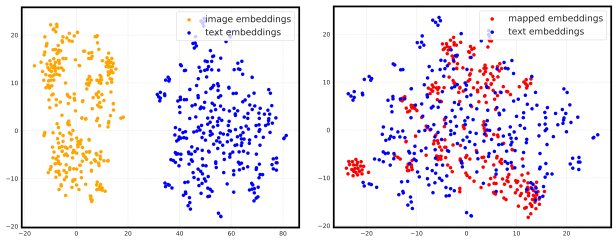



*Figure 3.* Analysis of the weighting factor $\alpha$. We set $\alpha$ from 0 to 1 on three datasets: (a) MS-COCO, (b) VOC2007, and (c) NUS-WIDE.




(a) Embeddings of image-text pairs from MS-COCO (b) Mapped embeddings and text embeddings

*Figure 4.* Visualization of embeddings in 2D space by t-SNE. We randomly sample 300 image-text pairs from the MSCOCO training set for visualization.

**Ablation of Fine-grained Image Embeddings.** Different from the existing works only focus on global image features, we leveraged both global and local image embeddings to enhance the ability to recognize multiple objects in the image. To verify the effectiveness of local image embeddings, we remove the local embedding and only use the global image embedding for classification. The experimental results are shown in Table 4. We can find that adding local image embeddings helps to improve the performance by a margin of 2.5%, 0.9%, and 0.6% on three datasets. The result demonstrates that the proposed fine-grained image embedding fusion approach is effective in enhancing the performance of MLR tasks.

### 4.4. Method Analysis

**Effect of Language Data Size.** To investigate the influence of language data size, we conducted experiments with different sizes (sentence numbers) of the generated text training data on the three datasets. The experimental results are provided in Figure 2. We find that increasing the size of training data can effectively improve the performance of our CoMC for all three datasets. It is worth noting that our method achieves impressive results when 50% of the language data is used. Overall, our method is data-efficient as the language data can be easily obtained.

**Effect of Different LLMs.** To investigate the influence of different text data generation models, we further used Llama-2-7B (Touvron et al., 2023) and ChatGLM-6B (Du et al., 2021) to generate the text dataset, both models are much smaller and more accessible than GPT-3. As shown in Table 5, we find that the model based on the text dataset generated by GPT-3 works best. This suggests that LLM with a larger number of parameters like GPT-3 can have a stronger capability to generate high-quality text data. It is worth noting that when we use smaller open-source models like Chatglm and Llama, the performance of our method is still significantly better than Zero-shot CLIP.

**Effect of Visual Backbones.** We evaluate the scalability of our model across three CLIP visual backbones: RN50, RN101, and ViT-B/32. Tabel 6 shows the result of CLIP, CLIP-DPT, and our proposed CoMC with different visual backbones. As the visual backbone network gets larger, the performance of our CoMC obtains consistent improvement. In addition, our model substantially outperforms CLIP-DPT with three different visual backbones.

**Analysis of the Weighting Factor $\alpha$.** We conduct experiments on the three datasets with different $\alpha$ between 0 and 1. The results are shown in Figure 3. The larger the $\alpha$ is, the larger the proportion of the original image embedding. When $\alpha$ is set to 0, no original image embedding is added. When $\alpha$ is set to 1, no cross-modal mapping is done and only image embedding is used for classification. On the three datasets, the model performance shows a similar trend when the value of $\alpha$ is increasing from 0 to 1. Fusing the original image embedding and mapped embedding achieves the best performance on all three datasets. When $\alpha$ is greater than the optimal value, the performance of the model shows a decreasing trend. This suggests that the modality gap does limit the performance of cross-modal transfer. When the value of $\alpha$ is less than 0.4, the performance of our model is significantly degraded on the three datasets. This is mainly because the text templates we set are too simple, using only the mapped embedding loses too much important visual information for MLR. These phenomena indicate that a balanced fusion of the original image embedding and mapped embedding is very important to obtain satisfactory performance.

### 4.5. Visualization

To further confirm the effectiveness of our cross-modal mapping, we randomly sample 300 image-text pairs from the MS-COCO training set and visualize their original CLIP embeddings and the mapped embeddings using t-SNE (Van der Maaten & Hinton, 2008). In Figure 4(a), we can see that image embeddings and text embeddings fall into two separate subspaces. A clear modality gap exists between images and texts. The mapped embeddings obtained by our method are shown in Figure 4(b), which demonstrates that our cross-modal mapping can effectively reduce the modality gap.

## 5. Conclusion

To mitigate the reliance on annotated image data, we propose a novel framework for zero-shot multi-label image recognition using language-only supervision. We utilize large language models (LLMs) to efficiently generate high-quality textual datasets for training. To address the modality gap between image and text, we design a simple yet effective cross-modal mapping method. This method enables our classifier, trained solely on language data, to be efficiently transferred to the visual modality for multi-label image recognition. Extensive experimental results demonstrate that, even without any image data for training, our method significantly outperforms existing zero-shot methods for multi-label recognition and achieves competitive performance compared to few-shot methods.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## Acknowledgements

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515030213), in part by the Shenzhen Higher Education Stability Support Program Project (Grant No. GXWD20220811173317002), and in part by National Natural Science Foundation of China (Grant No. 62372136).

## References

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7): 1425–1438, 2015.

Ben-Cohen, A., Zamir, N., Ben-Baruch, E., Friedman, I., and Zelnik-Manor, L. Semantic diversity learning for zero-shot multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 640–650, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *NeurIps*, volume 33, pp. 1877–1901, 2020.

Chen, T., Pu, T., Liu, L., Shi, Y., Yang, Z., and Lin, L. Heterogeneous semantic transfer for multi-label recognition with partial labels. *arXiv preprint arXiv:2205.11131*, 2022.

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., and Zheng, Y. 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*, pp. 1–9, 2009.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., and Tang, J. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

Guo, Z., Dong, B., Ji, Z., Bai, J., Guo, Y., and Zuo, W. Texts as images in prompt tuning for multi-label image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2808–2817, 2023.

He, S., Guo, T., Dai, T., Qiao, R., Shu, X., Ren, B., and Xia, S.-T. Open-vocabulary multi-label classification via multi-modal knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 808–816, 2023.

Huang, Z., Zeng, Z., Liu, B., Fu, D., and Fu, J. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

Huynh, D. and Elhamifar, E. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8776–8786, 2020.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Kim, J.-H., Jun, J., and Zhang, B.-T. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594. PMLR, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., and Zhou, M. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

Luo, H., Bao, J., Wu, Y., He, X., and Li, T. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pp. 23033–23044. PMLR, 2023.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pp. 529–544. Springer, 2022.

Naeem, M. F., Khan, M. G. Z. A., Xian, Y., Afzal, M. Z., Stricker, D., Van Gool, L., and Tombari, F. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15169–15179, 2023.

Narayan, S., Gupta, A., Khan, S., Khan, F. S., Shao, L., and Shah, M. Discriminative region-based multi-label zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8731–8740, 2021.

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G. S., and Dean, J. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

Novack, Z., McAuley, J., Lipton, Z. C., and Garg, S. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pp. 26342–26362. PMLR, 2023.

Pratt, S., Covert, I., Liu, R., and Farhadi, A. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15691–15701, 2023.

Pu, T., Sun, M., Wu, H., Chen, T., Tian, L., and Lin, L. Semantic representation and dependency learning for multi-label image recognition. *Neurocomputing*, 526:121–130, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Rahman, S., Khan, S., and Barnes, N. Deep0tag: Deep multiple instance learning for zero-shot image tagging. *IEEE Transactions on Multimedia*, 22(1):242–255, 2019.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Sun, X., Hu, P., and Saenko, K. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *arXiv preprint arXiv:2206.09541*, 2022.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Wang, L., Liu, Y., Du, P., Ding, Z., Liao, Y., Qi, Q., Chen, B., and Liu, S. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11186–11196, 2023.

Xu, S., Li, Y., Hsiao, J., Ho, C., and Qi, Z. A dual modality approach for (zero-shot) multi-label classification. *arXiv preprint arXiv:2208.09562*, 2022.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Zhang, T., He, S., Dai, T., Wang, Z., Chen, B., and Xia, S.-T. Vision-language pre-training with object contrastive learning for 3d scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7296–7304, 2024.

Zhang, Y., Gong, B., and Shah, M. Fast zero-shot image tagging. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5985–5994. IEEE, 2016.

Zhang, Y., HaoChen, J. Z., Huang, S.-C., Wang, K.-C., Zou, J., and Yeung, S. Diagnosing and rectifying vision models using language. In *The Eleventh International Conference on Learning Representations*, 2022.

Zhao, S., Zhang, Z., Schulter, S., Zhao, L., Vijay Kumar, B., Stathopoulos, A., Chandraker, M., and Metaxas, D. N. Exploiting unlabeled data with vision and language models for object detection. In *European Conference on Computer Vision*, pp. 159–175. Springer, 2022.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.