Super-Class Guided Transformer for Zero-Shot Attribute Classification

Sehyung Kim[*], **Chanhyeong Yang**[*], **Jihwan Park, Taehoon Song, Hyunwoo J. Kim**[†]
Department of Computer Science and Engineering, Korea University
{shkim129, 0814gerrardso, jseven7071, taehoons, hyunwoojkim}@korea.ac.kr

Abstract

Attribute classification is crucial for identifying specific characteristics within image regions. Vision-Language Models (VLMs) have been effective in zero-shot tasks by leveraging their general knowledge from large-scale datasets. Recent studies demonstrate that transformer-based models with class-wise queries can effectively address zero-shot multi-label classification. However, poor utilization of the relationship between seen and unseen attributes makes the model lack generalizability. Additionally, attribute classification generally involves many attributes, making maintaining the model's scalability difficult. To address these issues, we propose Super-class guided transFormer (SugaFormer), a novel framework that leverages super-classes to enhance scalability and generalizability for zero-shot attribute classification. SugaFormer employs Super-class Query Initialization (SQI) to reduce the number of queries, utilizing common semantic information from super-classes, and incorporates Multi-context Decoding (MD) to handle diverse visual cues. To strengthen generalizability, we introduce two knowledge transfer strategies that utilize VLMs. During training, Super-class guided Consistency Regularization (SCR) aligns model's features with VLMs using super-class guided prompts, and during inference, Zero-shot Retrieval-based Score Enhancement (ZRSE) refines predictions for unseen attributes. Extensive experiments demonstrate that SugaFormer achieves state-of-the-art performance across three widely-used attribute classification benchmarks under zero-shot, and cross-dataset transfer settings.

1 Introduction

Understanding an image at the region level is crucial for advancing computer vision. Building on the progress of image-level multi-label classification, region-level attribute classification takes this further by recognizing specific attributes within image regions. For example, it can describe a 'bus' as red, large, and parking, or the 'road' as wet and crowded. This task is vital for applications such as self-driving cars and image recommender systems, where the precise recognition of regional attributes directly impacts performance.

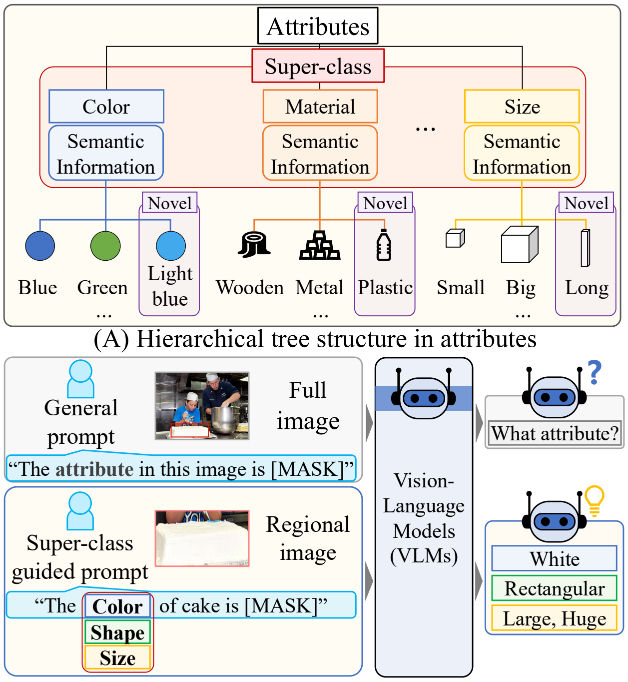


Figure 1: Hierarchical structure of attributes and effectiveness of super-class guided prompt. (A) Attributes belonging to the same super-class share common semantic information. (B) By leveraging super-class guided prompts, VLMs make more accurate predictions by distinguishing attributes within each super-class.

In recent years, the extension of visual recognition to zero-shot learning (Xian et al. 2018; Chen et al. 2023b), where models are required to predict novel classes not encountered during training, has achieved notable success (Zhou et al. 2023; Guo et al. 2024). A promising strategy is to leverage pre-trained Vision-Language Models (VLMs) (Radford et al. 2021; Li et al. 2023). This can be achieved through techniques such as fine-tuning VLMs on specific datasets, applying knowledge distillation (Yang et al. 2023; Liao et al. 2022), or directly integrating these models into the architecture (Gu et al. 2021; Ning et al.

---

[*] These authors contributed equally.
[†] Corresponding author.

2023). Given these insights, we leverage VLMs for zero-shot attribute classification, focusing on developing a effective and scalable approach to handle unseen classes.

One practical approach for zero-shot multi-label classification is a transformer-based classifier equipped with class-wise queries (Ridnik et al. 2023). This technique demonstrates strong performance in fine-grained classification and offers the flexibility to predict unseen labels. However, this line of work has two notable disadvantages. First, it lacks scalability for handling many classes, such as in attribute classification, since class-wise queries require substantial memory and computational cost as the number of classes increases. Second, the generalization of unseen classes is sub-optimal, mainly due to the poor utilization of the relationship between seen and unseen classes.

To this end, we propose **Su**per-class **g**uided tr**a**ns**Former** (**SugaFormer**), a framework that leverages super-classes to effectively utilize VLMs to address the challenges of zero-shot attribute classification. SugaFormer improves scalability and generalizability through Super-class Query Initialization (SQI), as shown in Fig. 1-(A), which illustrates the hierarchical relationship between super-classes and attribute classes. By leveraging this hierarchy, SQI reduces the number of queries and utilizes common semantic information by aligning attributes with their relevant super-classes. It also enhances performance with Multi-context Decoding (MD), which handles diverse visual cues. To strengthen generalizability, SugaFormer incorporates Super-class guided Consistency Regularization (SCR) during training, as illustrated in Fig. 1-(B), aligning its features with those of VLMs through the use of super-class guided prompt. During inference, it employs Zero-shot Retrieval-based Score Enhancement (ZRSE) to refine predictions for unseen attributes by integrating similarity scores from image and text embeddings.

The effectiveness of our SugaFormer framework is validated using three attribute classification datasets: VAW (Pham et al. 2021), LSA (Pham et al. 2022), and OVAD (Bravo et al. 2023). Our experiments and analyses demonstrate that SugaFormer achieves state-of-the-art performance in these benchmarks under zero-shot, and cross-dataset transfer settings by leveraging super-classes to accurately identify attributes and effectively VLMs.

In summary, our contributions are as follows:

- We present a Super-class guided Transformer (SugaFormer), a novel framework that improves scalability and generalizability for zero-shot attribute classification. Super-class Query Initialization (SQI) reduces the number of queries by utilizing super-classes and leveraging their common semantic information while Multi-context Decoding (MD) enhances performance by handling diverse visual cues.
- We propose two knowledge transfer strategies that leverage VLMs to enhance generalizability. Super-class guided Consistency Regularization (SCR) aligns its features with VLMs using super-class guided prompts during training. During inference, Zero-shot Retrieval-based Score Enhancement (ZRSE) refines predictions for unseen attributes by integrating similarity scores between image and text embeddings.
- Extensive experiments demonstrate that SugaFormer significantly improves scalability and generalizability, achieving state-of-the-art performance across three attribute classification benchmarks in zero-shot, and cross-dataset transfer settings.

2 Related Work

Attribute classification. Attribute classification aims to recognize attributes such as color, shape, etc., which belong to an object. Unlike other vision classification tasks like multi-label classification (Chua et al. 2009; Everingham et al. 2010; Irvin et al. 2019; Kuznetsova et al. 2018; Xu et al. 2022; Prokofiev and Sovrasov 2023; Liu et al. 2023b), attribute classification needs to consider multi-context visual cues. Depending on the attribute type, the model needs to determine which visual contexts to prioritize. Consequently, despite great success in previous studies (Liu et al. 2021; Ridnik et al. 2023) on multi-label classification, the necessity of a proper model for this task has arisen. Several works (Metwaly et al. 2022; Chen et al. 2023a) in this field have made much progress by leveraging various context information. For instance, GlideNet (Metwaly et al. 2022) employs different feature extractors for global, local, and intrinsic features and leverages the attention mechanism using a category estimator in a two-stage manner. Also, OvarNet (Chen et al. 2023a) exploits region proposals obtained from an RPN and utilizes extra caption data to fine-tune CLIP (Radford et al. 2021) model to exploit more context information for attribute classification.

VLMs and Zero-Shot Learning. Recently, VLMs (Radford et al. 2021; Li et al. 2023; Alayrac et al. 2022; Liu et al. 2023a; Yuan et al. 2021; Xiao et al. 2024; Zeng et al. 2023), pre-trained with large-scale image-text pairs have accomplished great advancements. Hence, a line of work (Cao et al. 2023; He et al. 2023) incorporates VLMs by transferring their general knowledge to downstream tasks and improves their performances. Therefore, in zero-shot attribute classification, OvarNet (Chen et al. 2023a) fine-tunes CLIP (Radford et al. 2021) for attributes with extra datasets and leverages its embeddings via prompt tuning. However, despite the improvements, the fine-tuning method in OvarNet may adversely affect the well-learned representations of the VLMs (Kumar et al. 2022). This could prevent the model from taking full advantage of the generalizability of the VLMs acquired during pretraining. Instead of the fine-tuning approach, we focus on the fact that a range of work (Kuo et al. 2023; Huang et al. 2024; Lin et al. 2022; Kuo et al. 2022; Kim et al. 2024) has shown that utilizing embeddings from frozen VLMs can achieve substantial performance on open-vocabulary settings across a variety of downstream tasks. Therefore, our work explores a method that leverages embeddings from frozen pre-trained VLMs.

3 Method

In this section, we introduce SugaFormer, a framework that leverages super-classes to enhance scalability and generalizability for zero-shot attribute classification. Before delving into SugaFormer, we briefly introduce the attribute classification task and base architecture in Sec. 3.1. We delineate the super-class query initialization and multi-context decoding of SugaFormer in Sec. 3.2. We then present knowledge transfer strategies to leverage the VLMs for improving generalization during training and inference in Sec. 3.3. The overall architecture of SugaFormer is illustrated in Fig. 2.

3.1 Preliminary

Problem setting. Attribute classification is the task of recognizing the positive attributes of an object in an image. Specifically, for each target object with bounding box $\mathbf{b} \in \mathbb{R}^4$, segmentation mask $\mathbf{m} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}}$, and object category $\mathbf{o}$ in an image $\mathbf{I} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 3}$, an attribute classification model predicts label vectors $\mathbf{Y} \in \{1, 0, -1\}^{\mathcal{N}_{\mathbf{a}}}$. Here, H and W represent the height and width of the image, $\mathcal{N}_{\mathbf{a}}$ is the number of attribute classes, and the values 1, 0, and -1 indicate positive, negative, and unknown attributes, respectively. In short, an attribute classifier maps a target object $(\mathbf{I}, \mathbf{b}, \mathbf{m}, \mathbf{o})$ to label vectors $\mathbf{Y}$.

In this work, we explore the zero-shot setting in attribute classification, where only a subset of attribute classes, *i.e.*, base classes $\mathbf{A}_{\text{base}}$, are used for training. During inference, all attribute classes $\mathbf{A} = \mathbf{A}_{\text{base}} \cup \mathbf{A}_{\text{novel}}$ are used to test the model's generalizability to unseen attribute classes.

Base architecture. We choose a transformer-based multi-label classifier ML-Decoder (Ridnik et al. 2023) as a base model due to its performance and simple architecture. Although this model has not been adopted for attribute classification, since this task can be seen as a multi-label classification for a target object, the model can be applied to attribution classification with minor modifications. For zero-shot learning, ML-Decoder uses word embeddings of attribute classes obtained from a language model (Devlin et al. 2019). Given i-th attribute class $\mathbf{a}_i \in \mathbf{A}$ and text encoder $\mathcal{T}_{\text{text}}$, the corresponding query $\mathbf{q}_i$ is initialized by its text embedding $\mathbf{t}_i$ as follows:

$$\mathbf{q}_i = \mathbf{t}_i = \mathcal{T}_{\text{text}}(\mathbf{a}_i). \tag{1}$$

Then, the query $\mathbf{q}_i$ is fed into the decoder:

$$\mathbf{q}'_i = \mathbf{CA}(\mathbf{q}_i, \mathbf{f}, \mathbf{f}) \quad \text{and} \quad \hat{\mathbf{q}}_i = \mathbf{FFN}(\mathbf{q}'_i), \tag{2}$$

where $\mathbf{f}$ is the visual feature map, $\mathbf{CA}(\cdot,\cdot,\cdot)$ is a cross-attention operation and $\mathbf{FFN}(\cdot)$ is a feed-forward operation. Finally, leveraging a output query $\hat{\mathbf{q}}$, the prediction score $\hat{p}$ is obtained as follows:

$$\hat{c}_i = \mathbf{t}_i \cdot \hat{\mathbf{q}}_i \quad \text{and} \quad \hat{p}_i = \sigma(\hat{c}_i), \tag{3}$$

where $\hat{c}_i$ is logit score for the i-th attribute class, and $\sigma(\cdot)$ is the sigmoid function. $\hat{p}_i$ is the prediction score for the i-th attribute class.

3.2 Super-class guided Transformer

We here delineate our **Su**per-class **g**uided tr**a**ns**Former** (**SugaFormer**) designed to enhance scalability and generalizability in zero-shot attribute classification. SugaFormer incorporates two key components: 1) super-class query initialization, a strategy that initializes decoder queries based on super-classes to utilize their common semantic information, and 2) multi-context decoding, a scheme that decodes queries with contextual feature maps to handle diverse visual cues.

Super-class Query Initialization. We propose super-class query initialization to enhance generalization capability and reduce the number of queries by leveraging the common semantic information within super-classes. Typically, previous methods for zero-shot multi-label classification (Ridnik et al. 2023) create one query per class. However, the class-wise query becomes inefficient with a high number of attribute classes (Pham et al. 2022). More importantly, class-wise queries are not effective for zero-shot attribute classification training, as a single class-wise query lacks semantic information between attributes and only receives supervised signals from its corresponding class. On the other hand, super-class queries capture the common semantic information shared by attributes within the corresponding super-class. Hence, for more effective training, we introduce super-class queries that are shared by a group of attribute classes. We initialize the super-class queries with the text embeddings of super-class names and visual representation of the target object for better query initialization. We adopt Q-Former in BLIP2 (Li et al. 2023) to extract the object-conditioned visual feature $\mathbf{v} \in \mathbb{R}^{d_q}$. The image $\mathbf{I}$ with the target object box $\mathbf{b}$ serves as the input of the visual backbone $\mathcal{V}_{\text{img}}$, which is connected to Q-Former. Then Q-Former extracts the output features $\hat{\mathbf{z}} \in \mathbb{R}^{\mathcal{N}_{\mathbf{z}}\times d_q}$ from the fixed set of learned queries $\mathbf{z} \in \mathbb{R}^{\mathcal{N}_{\mathbf{z}}\times d_q}$ by operating cross-attention with the feature maps from $\mathcal{V}_{\text{img}}$. We pool the $\hat{\mathbf{z}}$ to obtain the visual features. The above process is formulated as follows:

$$\begin{aligned} \hat{\mathbf{z}} &= \text{Q-Former}(\mathbf{z}, \mathcal{V}_{\text{img}}(\mathbf{\Phi}_{\text{crop}}(\mathbf{I}, \mathbf{b}))), \\ \mathbf{v} &= \mathbf{Pool}(\hat{\mathbf{z}}), \end{aligned} \tag{4}$$

where $\mathbf{\Phi}_{\text{crop}}$ represents the cropping function, and $\mathbf{Pool}(\cdot)$ denotes a pooling function, such as mean or max. Finally, we concatenate the j-th super-class query $\mathbf{q}_j$ with the object-conditioned visual features $\mathbf{v}$ to form a combined query $\tilde{\mathbf{q}}_j$, which is then projected onto a d_q-dimensional feature space:

$$\tilde{\mathbf{q}}_j = \text{Proj}([\mathbf{q}_j; \mathbf{v}]), \tag{5}$$

where $\text{Proj} \in \mathbb{R}^{d\times 2d_q}$ and $[\cdot;\cdot]$ are a linear projection and concatenation, respectively.

Multi-context Decoding. We propose a decoding strategy to utilize diverse contextual information to enhance performance. Our model first extracts three types of feature maps: cropped feature $\mathbf{f}_{\text{crop}}$, masked feature $\mathbf{f}_{\text{mask}}$, and image feature $\mathbf{f}_{\text{img}}$.

The cropped feature $\mathbf{f}_{\text{crop}}$ is extracted from the image cropped by bounding box $\mathbf{b}$ to capture the local information. We used a large-scale pretrained visual backbone $\mathcal{V}_{\text{img}}$, such as CLIP (Radford et al. 2021), as an encoder. To focus only on the target object (*i.e.*, foreground), the masked feature $\mathbf{f}_{\text{mask}}$ is extracted after removing the background context by ground-truth segmentation mask $\mathbf{m}$. Additionally, for incorporating the global context, the entire image is fed into the visual encoder to obtain $\mathbf{f}_{\text{img}}$. The detailed formulations

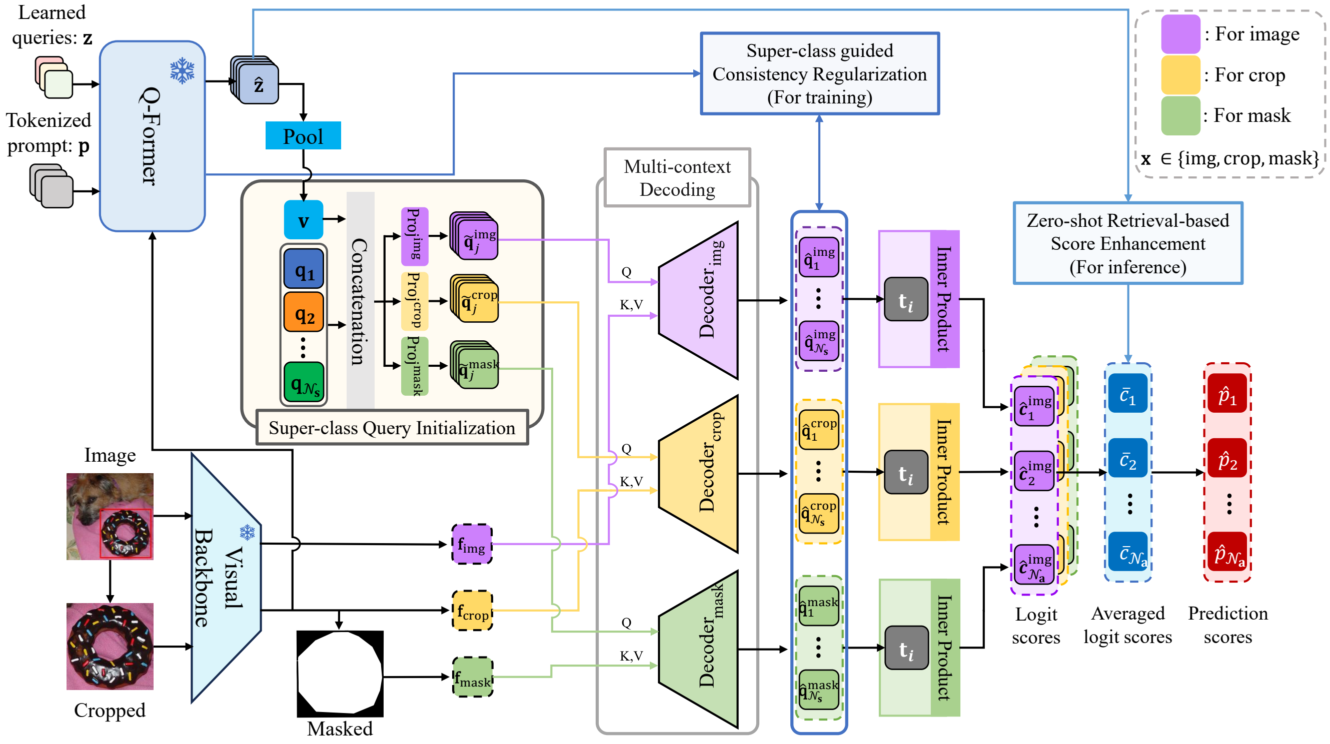


Figure 2: Model architecture. The overall pipeline of SugaFormer includes extracting multi-context visual features $\mathbf{f}_{\mathbf{x}}$ using an image, cropped image, and masked image. The super-class query $\mathbf{q}_j$ and pooled visual feature $\mathbf{v}$ are concatenated. The concatenated query passed through different projection layers $\text{Proj}^{\mathbf{x}}(\cdot)$, generating super-class queries $\tilde{\mathbf{q}}_j^{\mathbf{x}}$. Each $\text{Decoder}_{\mathbf{x}}(\cdot)$ processes its respective $\tilde{\mathbf{q}}_j^{\mathbf{x}}$ with corresponding visual feature maps $\mathbf{f}_{\mathbf{x}}$, producing output queries $\hat{\mathbf{q}}_j^{\mathbf{x}}$. Logit scores $\hat{c}_i^{\mathbf{x}}$ are computed via the inner product between $\mathbf{t}_i$ and $\hat{\mathbf{q}}_j^{\mathbf{x}}$. The averaged logit score $\bar{c}_i$ is used to calculate the prediction score $\hat{p}^i$ for the i-th attribute. To enhance generalizability, super-class guided consistency regularization is applied during training, and zero-shot retrieval-based score enhancement is used during inference. Best viewed in color.

for extracting each feature map are provided below.

$$\begin{aligned} \mathbf{f}_{\text{img}} \in \mathbb{R}^{hw\times d_v} &= \mathcal{V}_{\text{img}}(\mathbf{I}), \\ \mathbf{f}_{\text{crop}} \in \mathbb{R}^{hw\times d_v} &= \mathcal{V}_{\text{img}}(\mathbf{\Phi}_{\text{crop}}(\mathbf{I}, \mathbf{b})), \\ \mathbf{f}_{\text{mask}} \in \mathbb{R}^{hw\times d_v} &= \mathbf{f}_{\text{crop}} \odot \mathbf{\Phi}_{\text{resize}}(\mathbf{m}), \end{aligned} \tag{6}$$

where $\mathbf{\Phi}_{\text{crop}}$ and $\mathbf{\Phi}_{\text{resize}}$ are image cropping and resizing functions, respectively, and $\odot$ denotes element-wise multiplication. Similar to the query initialization in Eq. (1), a set of super-class queries $Q = \{\mathbf{q}_j\}_{j=1}^{\mathcal{N}_s}$ are initialized by $\mathbf{q}_j = \mathbf{t}_j = \mathcal{T}_{\text{text}}(\mathbf{s}_j)$, where $\mathbf{s}_j \in \mathbf{S}$ represents the j-th super-class in the super-class set $\mathbf{S} = \{\mathbf{s}_j\}_{j=1}^{\mathcal{N}_s}$. Then, multi-context decoding is performed using super-class queries Q and feature maps $\mathbf{f}_{\text{crop}}, \mathbf{f}_{\text{mask}}$, and $\mathbf{f}_{\text{img}}$. Here, each feature map is processed independently via cross-attention with queries Q for separate decoders $\text{Decoder}_{\mathbf{x}}(\cdot,\cdot,\cdot)$ to handle diverse visual cues, where $\mathbf{x} \in \{$ img, crop, mask $\}$. Also, for each decoder, the super-class queries Q are separately initialized using different linear projection $\text{Proj}^{\mathbf{x}}$ as follows:

$$\tilde{\mathbf{q}}_j^{\mathbf{x}} = \text{Proj}^{\mathbf{x}}([\mathbf{q}_j; \mathbf{v}]), \tag{7}$$

where $\tilde{\mathbf{q}}_j^{\mathbf{x}}$ represents the combined query for context $\mathbf{x}$. The output features of j-th super-class query for each decoder can be formulated as:

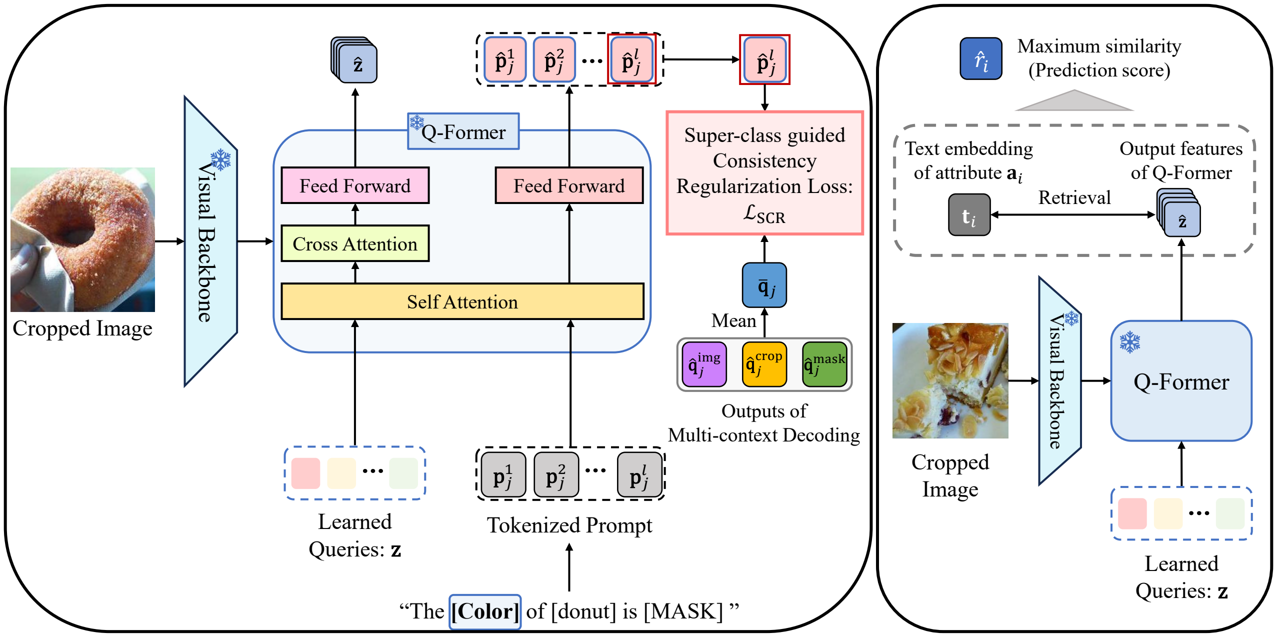
$$\begin{aligned} \hat{\mathbf{q}}_j^{\text{img}} &= \text{Decoder}_{\text{img}}(\tilde{\mathbf{q}}_j^{\text{img}}, \tilde{\mathbf{f}}_{\text{img}}, \tilde{\mathbf{f}}_{\text{img}}), \\ \hat{\mathbf{q}}_j^{\text{crop}} &= \text{Decoder}_{\text{crop}}(\tilde{\mathbf{q}}_j^{\text{crop}}, \tilde{\mathbf{f}}_{\text{crop}}, \tilde{\mathbf{f}}_{\text{crop}}), \\ \hat{\mathbf{q}}_j^{\text{mask}} &= \text{Decoder}_{\text{mask}}(\tilde{\mathbf{q}}_j^{\text{mask}}, \tilde{\mathbf{f}}_{\text{mask}}, \tilde{\mathbf{f}}_{\text{mask}}), \end{aligned} \tag{8}$$

where $\tilde{\mathbf{f}}_{\mathbf{x}} \in \mathbb{R}^{hw\times d}$ represents the projected features for each $\mathbf{x}$.

Attribute prediction. Our framework outputs the final attribute predictions by integrating all predictions from three decoders' outputs: $\hat{\mathbf{q}}_j^{\text{img}}, \hat{\mathbf{q}}_j^{\text{crop}}$, and $\hat{\mathbf{q}}_j^{\text{mask}}$. For context $\mathbf{x}$ and i-th attribute class, the logit score $\hat{c}_i^{\mathbf{x}}$ is computed by the the inner product between text embedding $\mathbf{t}_i$ and its corresponding super-class output query feature $\hat{\mathbf{q}}_j^{\mathbf{x}}$. Then, the final prediction score $\hat{p}_i$ for i-th attribute class is calculated as:

$$\hat{c}_i^{\mathbf{x}} = \mathbf{t}_i \cdot \hat{\mathbf{q}}_{j=\delta(i)}^{\mathbf{x}}, \text{ and } \hat{p}_i = \sigma(\bar{c}_i), \tag{9}$$

where $j = \delta(i)$ denotes the super-class index of the i-th attribute class and $\bar{c}_i = (\hat{c}_i^{\text{img}} + \hat{c}_i^{\text{crop}} + \hat{c}_i^{\text{mask}})/3$ denotes averaged logit score.



(A) Super-class guided Consistency Regularization (SCR)

(B) Zero-shot Retrieval-based Score Enhancement (ZRSE)

Figure 3: Illustration of knowledge transfer strategies. (A) During training, the Q-Former extracts a [MASK] token feature $\hat{\mathbf{p}}_j^l$ using a super-class guided prompt $\mathbf{p}_j$. This process leverages learned queries $\mathbf{z}$ and a tokenized prompt which is obtained by using a the prompt $\mathbf{p}_j$ that integrates the j-th super-class and the object class name. We compute $\mathcal{L}_{\text{SCR}}$ which is obtained by measuring L1 distance between the mean of output features $\bar{\mathbf{q}}_j$ from multi-context decoding and the [MASK] token feature $\hat{\mathbf{p}}_j^l$. (B) During inference, we employ ZRSE in which maximum similarity $\hat{r}_i$ obtained from Q-Former to compensate for novel classes. Note that we use the same frozen Q-Former.

3.3 Knowledge transfer strategies

In this section, we introduce knowledge transfer strategies by leveraging VLMs to improve generalizability for zero-shot attribute classification with 1) super-class guided consistency regularization and 2) zero-shot retrieval-based score enhancement.

Super-class guided Consistency Regularization. We here propose Super-class guided Consistency Regularization scheme using a Image-grounded Text Generation (ITG) with a super-class guided prompt as described in Fig. 3-(A). Given the cropped image of the target object as input, our approach uses Q-Former to generate texts by employing multimodal causal masked attention (Li et al. 2023) with the learned queries $\mathbf{z}$. The output feature corresponding to the prompt $\mathbf{p}$, which is tokenized into l tokens, is computed as:

$$\hat{\mathbf{p}} = \text{Q-Former}([\mathbf{z}; \mathbf{p}], \mathcal{V}_{\text{img}}(\mathbf{\Phi}_{\text{crop}}(\mathbf{I}, \mathbf{b}))), \tag{10}$$

where $\hat{\mathbf{p}} \in \mathbb{R}^{l \times d_q}$ denotes the features extracted from the prompt $\mathbf{p}$ including [MASK] token. Here, we construct a super-class guided prompt $\mathbf{p}_j$ for the j-th super-class as `'The [super-class] of the [object] is [MASK]'`. For instance, in Fig. 3-(A), we can use `'The [color] of the [donut] is [MASK]'` for super-class `[color]`. The super-class guided prompt $\mathbf{p}_j$ constructed with j-th super-class name encourages Q-Former to generate text tokens more relevant to attribute classification. Instead of the generated text, we use the output feature $\hat{\mathbf{p}}_j^l \in \mathbb{R}^{d_q}$, where the l-th token corresponds to the [MASK] token from the prompt $\mathbf{p}_j$. We compute the L1 loss between the output feature $\hat{\mathbf{p}}_j^l$ and the averaged output features $\bar{\mathbf{q}}_j = (\hat{\mathbf{q}}_j^{\text{img}} + \hat{\mathbf{q}}_j^{\text{crop}} + \hat{\mathbf{q}}_j^{\text{mask}})/3$ obtained from multiple decoders $\text{Decoder}_{\mathbf{x}}$. The loss function for super-class guided consistency regularization $\mathcal{L}_{\text{SCR}}$ is formulated as follows:

$$\hat{\mathbf{p}}_j = \text{Q-Former}([\mathbf{z}; \mathbf{p}_j], \mathcal{V}_{\text{img}}(\mathbf{\Phi}_{\text{crop}}(\mathbf{I}, \mathbf{b}))), \tag{11}$$

$$\mathcal{L}_{\text{SCR}} = \sum_{j=1}^{\mathcal{N}_s} |\hat{\mathbf{p}}_j^l - \bar{\mathbf{q}}_j|. \tag{12}$$

The total loss function of ours can be written as:

$$\mathcal{L}_{\text{total}} = \sum_{\mathbf{x}} \mathcal{L}_{\text{asym}}^{\mathbf{x}} + \lambda \mathcal{L}_{\text{SCR}}, \tag{13}$$

where $\mathcal{L}_{\text{asym}}$ is an asymmetric loss function for multi-label classification (Ridnik et al. 2021), and λ denotes the weight for $\mathcal{L}_{\text{SCR}}$.

Zero-shot Retrieval-based Score Enhancement. We propose a new plug-and-play module to enhance the final predictions on unseen attribute classes by leveraging the prediction scores from Q-Former, as described in Fig. 3-(B).

Method	VLM	Attribute-related data for training	$\mathbf{AP_{base}}$	$\mathbf{AP_{novel}}$	$\mathbf{AP_{all}}$
CLIP (Radford et al. 2021)	CLIP	–	50.04	46.54	49.60
BLIP2 (Li et al. 2023)	BLIP2	–	47.50	46.02	47.31
CLIP-Attr (Chen et al. 2023a)	CLIP	VAW_{base}	67.90	57.39	66.92
CLIP-Attr (Chen et al. 2023a)	CLIP	VAW_{base} + CC-3M-sub	69.79	59.16	68.87
CLIP-Attr (Chen et al. 2023a)	CLIP	VAW_{base} + CC-3M-sub + COCO-Cap-sub	70.24	57.73	69.03
OvarNet (Chen et al. 2023a)	CLIP	VAW_{base}	68.27	53.75	66.85
OvarNet (Chen et al. 2023a)	CLIP	VAW_{base} + CC-3M-sub	69.30	55.44	67.96
OvarNet (Chen et al. 2023a)	CLIP	VAW_{base} + CC-3M-sub + COCO-Cap-sub	69.80	56.4	68.52
ML-Decoder† (Ridnik et al. 2023)	BLIP2	VAW_{base}	73.12	53.38	70.61
SugaFormer	BLIP2	VAW_{base}	**75.18**	**60.59**	**73.32**

Table 1: Results on the VAW in zero-shot setting. † denotes the results obtained from our implementation. CC-3M-sub (Changpinyo et al. 2021) and COCO-Cap-sub (Lin et al. 2014) represent additional caption datasets filtered to extract attribute-related information for training.

In addition to the classification score for the i-th attribute class $\bar{c}_i$, obtained by averaging the logit scores from the multi-context decoder, we compute the attribute classification score using Q-Former, following a process similar to Image-Text Contrastive (ITC). In Q-Former, each output embedding $\hat{\mathbf{z}}$ generated from the learned queries $\mathbf{z}$ is individually compared to the text embedding, and the maximum similarity score is selected as the overall similarity score between the given image and text. Similar to the above process, we first calculate the text embedding of i-th attribute class $\mathbf{t}_i$, and each k-th output embedding of Q-Former $\hat{\mathbf{z}}_k$, and use maximum similarity $\hat{r}_i$ as a prediction score of i-th attribute class from Q-Former. This process can be formulated as:

$$\hat{r}_i = \max_{k=1,2,\ldots,\mathcal{N}_{\mathbf{z}}} \mathbf{t}_i \cdot \hat{\mathbf{z}}_k^{\top}. \tag{14}$$

Finally, we choose Top-K scores from $\mathbf{R} = \{\hat{r}_i\}_{i=1}^{\mathcal{N}_{\mathbf{a}}}$ and add it to the corresponding predictions of attribute classes from our model. Then the Eq. (9) can be re-written as:

$$\hat{p}_i = \begin{cases} \sigma(\bar{c}_i + \hat{r}_i), & \text{if } i \in \text{Top-K}(\mathbf{R}) \\ \sigma(\bar{c}_i), & \text{otherwise} \end{cases} \tag{15}$$

where Top-K($\mathbf{R}$) refers to the list of attribute class indexes with the Top-K similarity scores from Q-Former.

4 Experiments

Here, we evaluate SugaFormer using three attribute classification benchmark datasets: VAW (Pham et al. 2021), LSA (Pham et al. 2022) in the zero-shot setting, and OVAD (Bravo et al. 2023) in the cross-dataset setting. We demonstrate that SugaFormer effectively learns the semantic information shared by attributes within the same super-class, enhancing scalability and generalizability.

4.1 Datasets

VAW. VAW (Pham et al. 2021) consists of 620 attribute classes with object instances, including segmentation masks, box coordinates, and class names. The dataset combines images from VGPhraseCut (Wu et al. 2020) and GQA (Hudson and Manning 2019). For zero-shot attribute classification, following prior work (Chen et al. 2023a), we use half of the 'tail' attributes and 15% of 'medium' attributes as novel classes, resulting in 79 novel classes and 541 base classes. We adopt eight predefined super-classes: color, material, shape, size, texture, state, action, other.

LSA. LSA (Pham et al. 2022) aggregates images and attributes from various datasets, including Visual Genome, GQA, COCO-Attributes, Flickr30K-Entities, and MS-COCO. We evaluate under zero-shot setting (common-to-rare), with 5526 base attributes and 4012 novel attributes.

OVAD. OVAD (Bravo et al. 2023) introduces a test-only cross-dataset benchmark featuring 117 attributes and object instances for open-vocabulary attribute detection. The dataset organizes attributes into three subsets: 'head' (15 frequently occurring attributes), 'medium' (53 moderately frequent attributes), and 'tail' (49 rare attributes) based on thresholds defined by attribute annotation frequency. Experiments assume ground-truth boxes are given during testing.

4.2 Evaluation metric

Our evaluation metric is average precision (AP), assessed across three categories: base classes (AP_{base}), novel classes (AP_{novel}), and all attribute classes (AP_{all}). AP_{base} indicates the precision for "seen" labels present during training, reflecting its ability to recognize learned attributes. AP_{novel} indicates the precision for "unseen" labels, which were not encountered during training. AP_{all} provides an precision of the model's performance across both seen and unseen labels.

4.3 Implementation details

Our model employs ViT-g/14 as its visual backbone. We keep all parameters within the visual backbone during training and Q-Former frozen. We establish a baseline using ML-Decoder (Ridnik et al. 2023) with the same visual backbone to ensure a fair comparison. All training and evaluations are conducted on NVIDIA RTX 3090. Further implementation

Method	Num. Query	AP$_{\text{novel}}$
CLIP (Radford et al. 2021)	-	2.63
BLIP2 (Li et al. 2023)	-	2.58
TAP (Pham et al. 2022)	-	5.37
OvarNet (Chen et al. 2023a)	-	5.48
ML-Decoder (Ridnik et al. 2023)	9538	out of memory
SugaFormer	9	**5.80**

Table 2: Results on the LSA in the zero-shot setting.

Method	AP$_{\text{all}}$	AP$_{\text{head}}$	AP$_{\text{medium}}$	AP$_{\text{tail}}$
ALBEF (Li et al. 2021)	21.0	44.2	23.9	9.4
BLIP (Li et al. 2022)	24.3	51.0	28.5	9.7
BLIP2 (Li et al. 2023)	25.5	49.7	30.4	10.8
X-VLM (Zeng, Zhang, and Li 2022)	28.1	49.7	34.2	**12.9**
OVAD (Bravo et al. 2023)	21.4	48.0	26.9	5.2
CLIP-Attr (Chen et al. 2023a)	26.1	55.0	31.9	8.5
OvarNet (Chen et al. 2023a)	28.6	58.6	35.5	9.5
SugaFormer	**29.7**	**58.8**	**36.6**	10.8

Table 3: Results on the OVAD in the cross-dataset transfer setting.

details, including hyperparameters, can be found in the supplementary materials.

4.4 State-of-the-art comparison

Zero-shot attribute classfication. To validate the effectiveness of SugaFormer, we compare it with state-of-the-art methods and our baseline. In Tab.1, SugaFormer achieves an AP$_{\text{novel}}$ of 60.59 in the VAW zero-shot setting, demonstrating the benefits of leveraging super-classes to enhance generalizability for novel attributes. In Tab.2, SugaFormer achieves an AP$_{\text{novel}}$ of 5.80 in the LSA zero-shot setting, outperforming other methods. We could not train the baseline due to out-of-memory issues caused by the many queries. These results highlight SugaFormer's superior performance in zero-shot attribute classification.

Cross-dataset transfer. In Tab. 3, we compare SugaFormer with previous methods on the OVAD in the cross-dataset transfer setting. Our model is trained on the VAW dataset. It is worth mentioning that training on VAW performs poorly compared to image-caption pair datasets in cross-dataset transfer, as discussed in OVAD (Bravo et al. 2023). Nevertheless, SugaFormer achieves an AP$_{\text{all}}$ of 29.7, surpassing the previous best method by approximately across all categories.

4.5 Ablation study and analysis

SQI	MD	SCR	ZRSE	AP$_{\text{base}}$	AP$_{\text{novel}}$	AP$_{\text{all}}$
ML-Decoder (baseline)				73.12	53.38	70.61
✓				72.86	55.90	70.67
✓	✓			74.38	56.25	72.07
✓	✓	✓		74.49	58.38	72.44
✓	✓	✓	✓	**75.18** (+2.06)	**60.59** (+7.21)	**73.32** (+2.71)

Table 4: Ablations on key components.

Prompt type	AP$_{\text{base}}$	AP$_{\text{novel}}$	AP$_{\text{all}}$
no prompt	75.06	58.85	73.00
general prompt	73.84	55.64	71.52
super-class guided prompt	**75.18**	**60.59**	**73.32**

Table 5: Analysis on regularization strategies.

Ablation on key components. In Tab. 4, we conducted ablation studies on the VAW in the zero-shot setting by gradually incorporating suggested methods into our baseline model. When incorporating techniques, such as super-class query initialization (SQI), super-class guided consistency regularization (SCR), and zero-shot retrieval-based score enhancement (ZRSE), We observed improvements in AP$_{\text{novel}}$ by approximately 2.52, 2.13, and 2.21, respectively. Moreover, multi-context decoding (MD) boosts AP$_{\text{all}}$, with an improvement of 1.40. When all components were combined, the model achieved its highest performance, outperforming the baseline by 2.06 for AP$_{\text{base}}$, 7.21 for AP$_{\text{novel}}$, and 2.71 AP$_{\text{all}}$. These results demonstrate that leveraging super-classes enhances the generalizability of attribute classification while utilizing diverse visual cues improves overall performance.

Analysis on regularization strategies. In this experiment, we explore three regularization strategies using different prompts on the VAW in the zero-shot setting. First, the no prompt approach used only image features without any textual prompt. A caption was generated through the ITG process in Q-Former, and the resulting embedding was equally compared to all super-class queries. Second, the general prompt without a super-class followed the format '`The attribute of the [object] is [MASK].`' using the `[MASK]` token for consistency. Finally, the super-class guided prompt employed the format '`The [super-class] of the [object] is [MASK].`' ensuring consistency through the use of the [MASK] token. As shown in Tab. 5, the best performance was achieved when prompts were tailored to each super-class, demonstrating that utilizing super-classes in prompts effectively extracts features particularly useful for attribute prediction.

5 Conclusion

We propose SugaFormer, a Super-class guided Transformer framework designed to address scalability and generalizability in zero-shot attribute classification. SugaFormer employs Super-class Query Initialization (SQI) to reduce queries by utilizing shared semantic information and Multi-context Decoding (MD) to enhance performance using diverse visual cues. Additionally, we introduce two knowledge transfer strategies leveraging VLMs: Super-class guided Consistency Regularization (SCR) aligns features during training, and Zero-shot Retrieval-based Score Enhancement (ZRSE) refines predictions during inference. Experiments on three benchmarks show that SugaFormer outperforms existing methods across zero-shot, and cross-dataset transfer settings.

Acknowledgements

This work was partly supported by the Institute of Information & Coummunications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIP & MSIT) (No. RS-2024-00443251, 40%, Accurate and Safe Multimodal, Multilingual Personalized AI Tutors; No. RS-2024-00457882, 30%, National AI Research Lab Project; No. IITP-2024-RS-2024-00436857, 30%, Information Technology Research Center ITRC).

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Bravo, M. A.; Mittal, S.; Ging, S.; and Brox, T. 2023. Open-vocabulary attribute detection. In *CVPR*.

Cao, Y.; Tang, Q.; Su, X.; Song, C.; You, S.; Lu, X.; and Xu, C. 2023. Detecting Any Human-Object Interaction Relationship: Universal HOI Detector with Spatial Prompt Learning on Foundation Models. In *NeurIPS*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Chen, K.; Jiang, X.; Hu, Y.; Tang, X.; Gao, Y.; Chen, J.; and Xie, W. 2023a. Ovarnet: Towards open-vocabulary object attribute recognition. In *CVPR*.

Chen, Z.; Huang, Y.; Chen, J.; Geng, Y.; Zhang, W.; Fang, Y.; Pan, J. Z.; and Chen, H. 2023b. Duet: Cross-modal semantic grounding for contrastive zero-shot learning. In *AAAI*.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *ACM*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *IJCV*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv*.

Guo, Y.; Liu, Y.; Li, J.; Wang, W.; and Jia, Q. 2024. Unseen No More: Unlocking the Potential of CLIP for Generative Zero-shot HOI Detection. In *ACM*.

He, S.; Guo, T.; Dai, T.; Qiao, R.; Shu, X.; Ren, B.; and Xia, S.-T. 2023. Open-vocabulary multi-label classification via multi-modal knowledge transfer. In *AAAI*.

Huang, X.; Zhou, H.; Yao, K.; and Han, K. 2024. FROSTER: Frozen CLIP Is A Strong Teacher for Open-Vocabulary Action Recognition. In *ICLR*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*.

Kim, J.; Cho, E.; Kim, S.; and Kim, H. J. 2024. Retrieval-Augmented Open-Vocabulary Object Detection. In *CVPR*.

Kumar, A.; Raghunathan, A.; Jones, R. M.; Ma, T.; and Liang, P. 2022. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. In *ICLR*.

Kuo, W.; Cui, Y.; Gu, X.; Piergiovanni, A.; and Angelova, A. 2022. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv*.

Kuo, W.; Cui, Y.; Gu, X.; Piergiovanni, A.; and Angelova, A. 2023. F-VLM: Open-Vocabulary Object Detection upon Frozen Vision and Language Models. In *ICLR*.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J. R. R.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Duerig, T.; and Ferrari, V. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *CoRR*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Liao, Y.; Zhang, A.; Lu, M.; Wang, Y.; Li, X.; and Liu, S. 2022. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Lin, Z.; Geng, S.; Zhang, R.; Gao, P.; de Melo, G.; Wang, X.; Dai, J.; Qiao, Y.; and Li, H. 2022. Frozen CLIP Models are Efficient Video Learners. In *ECCV*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. In *NeurIPS*.

Liu, R.; Huang, J.; Li, T. H.; and Li, G. 2023b. Causality compensated attention for contextual biased visual recognition. In *ICLR*.

Liu, S.; Zhang, L.; Yang, X.; Su, H.; and Zhu, J. 2021. Query2label: A simple transformer way to multi-label classification. *arXiv*.

Metwaly, K.; Kim, A.; Branson, E.; and Monga, V. 2022. Glidenet: Global, local and intrinsic based dense embedding network for multi-category attributes prediction. In *CVPR*.

Ning, S.; Qiu, L.; Liu, Y.; and He, X. 2023. HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models. In *CVPR*.

Pham, K.; Kafle, K.; Lin, Z.; Ding, Z.; Cohen, S.; Tran, Q.; and Shrivastava, A. 2021. Learning to predict visual attributes in the wild. In *CVPR*.

Pham, K.; Kafle, K.; Lin, Z.; Ding, Z.; Cohen, S.; Tran, Q.; and Shrivastava, A. 2022. Improving closed and open-vocabulary attribute prediction using transformers. In *ECCV*.

Prokofiev, K.; and Sovrasov, V. 2023. Combining Metric Learning and Attention Heads for Accurate and Efficient Multilabel Image Classification. In *VISIGRAPP*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric loss for multi-label classification. In *ICCV*.

Ridnik, T.; Sharir, G.; Ben-Cohen, A.; Ben-Baruch, E.; and Noy, A. 2023. ML-Decoder: Scalable and versatile classification head. In *WACV*.

Wu, C.; Lin, Z.; Cohen, S.; Bui, T.; and Maji, S. 2020. Phrasecut: Language-based image segmentation in the wild. In *CVPR*.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE*.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*.

Xu, S.; Li, Y.; Hsiao, J.; Ho, C.; and Qi, Z. 2022. Open vocabulary multi-label classification with dual-modal decoder on aligned visual-textual features. *arXiv*.

Yang, P.; Xie, M.-K.; Zong, C.-C.; Feng, L.; Niu, G.; Sugiyama, M.; and Huang, S.-J. 2023. Multi-Label Knowledge Distillation. In *ICCV*.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A new foundation model for computer vision. *arXiv*.

Zeng, Y.; Zhang, X.; and Li, H. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *ICML*.

Zeng, Y.; Zhang, X.; Li, H.; Wang, J.; Zhang, J.; and Zhou, W. 2023. X 2-vlm: All-in-one pre-trained model for vision-language tasks. *IEEE*.

Zhou, Z.; Zhang, B.; Lei, Y.; Liu, L.; and Liu, Y. 2023. ZegCLIP: Towards Adapting CLIP for Zero-shot Semantic Segmentation. In *CVPR*.