Leveraging MLLM Embeddings and Attribute Smoothing for Compositional Zero-Shot Learning

Xudong Yan[1], **Songhe Feng**[1*], **Yang Zhang**[1], **Jian Yang**[2], **Yueguan Lin**[2], **Haojun Fei**[2*]
[1]School of Computer Science and Technology, Beijing Jiaotong University
[2]Qifu Technology
{xud_yan, shfeng, chefzhang}@bjtu.edu.cn, {yangjian1, linyueguan, feihaojun}-jk@360shuke.com

Abstract

Compositional zero-shot learning (CZSL) aims to recognize novel compositions of attributes and objects learned from seen compositions. Previous works disentangle attributes and objects by extracting shared and exclusive parts between the image pair sharing the same attribute (object), as well as aligning them with pretrained word embeddings to improve unseen attribute-object recognition. Despite the significant achievements of existing efforts, they are hampered by three limitations: (1) The efficacy of disentanglement is compromised due to the influence of the background and the intricate entanglement of attributes with objects in the same parts. (2) Existing word embeddings fail to capture complex multimodal semantic information. (3) Overconfidence exhibited by existing models in seen compositions hinders their generalization to novel compositions. Being aware of these, we propose a novel framework named multimodal large language model (MLLM) embeddings and attribute smoothing guided disentanglement for CZSL. First, we leverage feature adaptive aggregation modules to mitigate the impact of background, and utilize learnable condition masks to capture multi-granularity features for disentanglement. Moreover, the last hidden states of MLLM are employed as word embeddings for their superior representation capabilities. Furthermore, we propose attribute smoothing with auxiliary attributes generated by the large language model (LLM) for seen compositions to address the overconfidence challenge. Extensive experiments demonstrate that our method achieves state-of-the-art performance on three challenging datasets. The supplementary material and source code will be available at https://github.com/xud-yan/Trident.

1 Introduction

As for the study of compositional generalization ability inherent in human beings, compositional zero-shot learning (CZSL) [Misra *et al.*, 2017; Nagarajan and Grauman, 2018; Purushwalkam *et al.*, 2019] is proposed to enable machines to recognize unseen compositions by leveraging knowledge of attributes and objects (*i.e.*, primitives) learned from seen compositions. Specifically, in the training phase, the models are provided with images of seen compositions (*e.g.*, `ripe orange` and `peeled apple`). During the testing phase, given an image that depicts a novel composition (*e.g.*, `peeled orange`), models are assigned to correctly recognize it [Zhang *et al.*, 2022].

Prior works [Misra *et al.*, 2017; Nan *et al.*, 2019; Naeem *et al.*, 2021] focus on mapping the visual features and the word embeddings of compositions into a joint space. These methods have poor generalization capabilities on unseen compositions due to the recombination of primitives. Therefore, recent studies [Saini *et al.*, 2022; Wang *et al.*, 2023; Li *et al.*, 2024; Zhang *et al.*, 2024] consider visual disentanglement. Among them, some prominent works [Hao *et al.*, 2023] deploy a triplet of images to disentangle visual features: a given image and two supplementary images, each sharing either the same attribute or object as the given image. The triplet of images is treated as two image pairs for subsequent analysis. These approaches aim to disentangle attribute and object by extracting the shared and exclusive features of the image pair, as well as aligning them with word embeddings (*e.g.*, GloVe [Pennington *et al.*, 2014]), as shown in Figure 1. Although these pioneering research studies have made great progress, they exhibit three limitations:

L1: Disentanglement is impeded due to the influence of the background and the intricate entanglement between attributes and objects in the same parts of images. On the one hand, models tend to extract the background features unique to one image in the pair as the disentangled exclusive features. On the other hand, some existing methods [Ruis *et al.*, 2021; Saini *et al.*, 2022] compute the similarity of the image pair for disentanglement at the spatial level. However, this paradigm is limited by the frequent entanglement of attributes and objects within the same image regions.

L2: Existing word embeddings lack the depth needed to capture complex multimodal semantic information. To begin with, word embeddings (*e.g.*, GloVe [Pennington *et al.*, 2014]) are grounded in word frequency and contextual co-occurrence, overlooking the high-level semantic nuances [Sarzynska-Wawer *et al.*, 2021]. Moreover, the pro-

*Corresponding authors

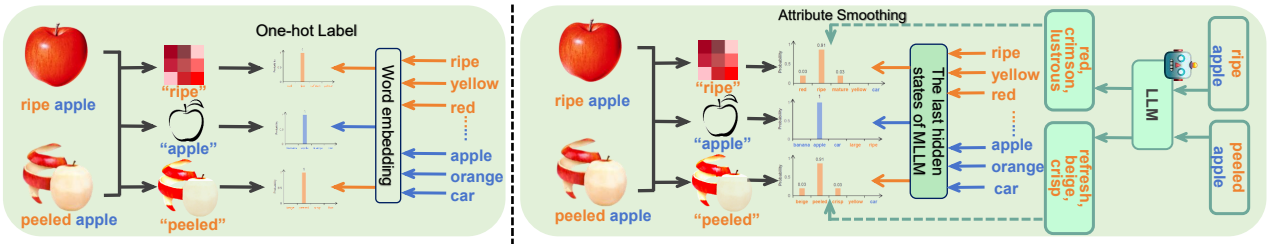

Figure 1: A general comparison between the existing method and our proposed **TRIDENT**. Note that, we only present the representation learning of an image pair sharing the object for brevity.

cess of aligning visual features with word embeddings can be viewed as a form of cross-modal matching; however, these word embeddings are trained only in a single text modality, failing to capture multimodal information between images and texts.

L3: Existing methods display excessive confidence in seen compositions, impairing their ability to generalize towards novel compositions. Specifically, due to the one-hot label used during training, these approaches are limited by learning only one disentangled attribute and object, neglecting the fact that objects naturally exhibit multiple attributes [Xu *et al.*, 2024]. Consequently, models exhibit overconfidence in the disentangled ground-truth attribute, treating other attributes that can describe the object as negative ones, which results in the diminished generalization to unseen compositions.

Being aware of these, we propose a novel framework named multimodal large language model (MLLM) embeddings and at**TR**ibute smooth**I**ng gui**DE**d dise**NT**anglement (**TRIDENT**), which consists of three major modules: visual feature extraction, attribute-object disentanglement, and feature alignment. The first module leverages feature adaptive aggregation (FAA) modules to mitigate the impact of background noise, and exploits learnable condition masks for multi-granularity feature learning at the dimensional level to improve subsequent disentanglement. The second module aims at leveraging shared and exclusive weights of image pairs to disentangle attributes and objects under the paradigm that apart from the shared features of the image pair, each image has its own exclusive features. The third module is intended to align the visual features of compositions and disentangled primitives with the last hidden states of MLLM (*i.e.*, MLLM embeddings). This is inspired by prior works [Wang and Kuo, 2020; Muennighoff, 2022; Muennighoff *et al.*, 2024], which reveal that the last hidden states of (M)LLM exhibit powerful representation capabilities in embedding tasks (*e.g.*, retrieval and classification). Moreover, to tackle the issue that the overconfidence of the models regarding the ground-truth attribute hinders them from generalizing to unseen compositions, we exploit the large language model (LLM) to generate auxiliary attributes for compositions and perform label smoothing for attributes (*i.e.*, attribute smoothing).

In summary, the contributions of our work are three-fold:

1. We propose novel feature adaptive aggregation modules to reduce the impact of background, and utilize learnable condition masks to capture multi-granularity features at the dimensional level for disentanglement in CZSL.
2. We employ both LLM and MLLM to guide attribute-object disentanglement by generating auxiliary attributes and representing primitive words for CZSL, respectively.
3. Extensive experiments conducted on three challenging datasets (MIT-States [Isola *et al.*, 2015], C-GQA [Naeem *et al.*, 2021], and VAW-CZSL [Saini *et al.*, 2022]) show that **TRIDENT** has achieved state-of-the-art performance.

2 Related Work

Compositional zero-shot learning (CZSL). Prior works in CZSL can be broadly divided into two main streams. One main stream is learning visual representations and textual labels of compositions in a joint space. SymNet [Li *et al.*, 2020] aims to learn the symmetry property in compositions. Co-CGE [Mancini *et al.*, 2022] leverages a graph convolutional neural network to learn composition representations. The other main stream aims at disentangling visual representations of primitives to reduce composition learning to primitive learning. SCEN [Li *et al.*, 2022] leverages contrastive loss to excavate discriminative prototypes of primitives. CANet [Wang *et al.*, 2023] learns the conditional attribute conditioned on the recognized object and the input image. More recent works [Nayak *et al.*, 2023; Lu *et al.*, 2023; Huang *et al.*, 2024] leverage the encyclopedic knowledge of pre-trained vision-language models (VLM) like CLIP [Radford *et al.*, 2021] to encode and align images and texts.

Large language model (LLM). LLMs have realized significant advancements thanks to the scaling up of training data and the increase in the number of parameters. Early models, such as GPT-2 [Radford *et al.*, 2019], initially exhibit strong capabilities in understanding and generating human-like language. Subsequently, GPT-3 [Brown *et al.*, 2020] and LLaMA [Touvron *et al.*, 2023] demonstrate great breakthroughs across numerous language benchmarks.

Expanding on LLMs, multimodal large language models (MLLM) incorporate a visual encoder for vision-language tasks. Flamingo [Alayrac *et al.*, 2022] integrates Vision Transformer (ViT) [Dosovitskiy *et al.*, 2021] and LLM by gated cross-attention. LLaVA [Liu *et al.*, 2024b] and LLaVA v1.5 [Liu *et al.*, 2024a] introduce visual instruction tuning to enhance the instruction following capability. The visual understanding part of LLaVA v1.5 consists of a ViT and an MLP cross-modal connector. We choose LLaVA v1.5 as our foundational MLLM for its state-of-the-art performance.

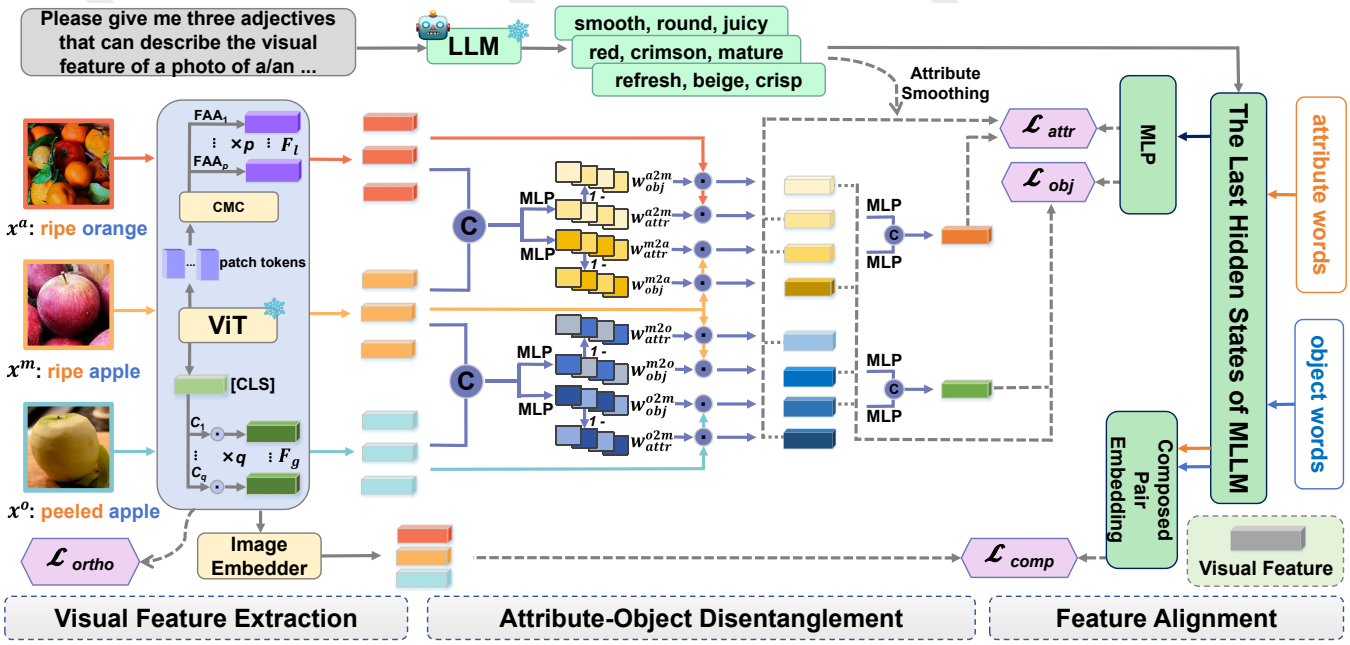

Figure 2: The overall architecture of our proposed **TRIDENT**. The model consists of three major modules: (a) visual feature extraction, (b) attribute-object disentanglement, and (c) feature alignment.

Recently, exploring the powerful embedding capabilities of (M)LLM to handle representation tasks (*e.g.*, retrieval) has emerged as a prominent research domain. SGPT [Muennighoff, 2022] exploits the last hidden states of LLM for the input token sequence or a special learnable token to derive representational embeddings. Subsequently, GritLM [Muennighoff *et al.*, 2024] applies mean pooling over the last hidden states of LLM to produce the textual embeddings.

3 Approach

3.1 Task Formulation

Compositional zero-shot learning (CZSL) aims at learning a model that can recognize unseen compositions of attributes and objects that are learned from seen compositions. Given an attribute set $\mathcal{A}$ and an object set $\mathcal{O}$, the attributes and objects are composed to form a composition set $\mathcal{C} = \mathcal{A} \times \mathcal{O}$. The composition set $\mathcal{C}$ is divided into two disjoint sets: the seen composition set $\mathcal{C}_s$ and the unseen composition set $\mathcal{C}_u$, *i.e.*, $\mathcal{C}_s \cap \mathcal{C}_u = \varnothing$. The model learns from a seen training set $\mathcal{D}_{tr} = \{(x_s, c_s)\}$, where x_s is an image of the seen composition label $c_s \in C_s$. Following the Generalized CZSL [Purushwalkam *et al.*, 2019], the model is evaluated on a predefined test set $\mathcal{D}_{te} = \{(x_{te}, c_{te})\}$, where x_{te} is a test image from the predefined composition subset $\mathcal{C}_{te}$ of $\mathcal{C}$, *i.e.*, $\mathcal{C}_{te} \subseteq \mathcal{C}$, and $c_{te} \in \mathcal{C}_{te}$ is the label of x_{te}.

3.2 TRIDENT

As the major novelty, we propose a novel framework named MLLM embeddings and attribute smoothing guided disentanglement (**TRIDENT**), as shown in Figure 2. It consists of three modules: (1) visual feature extraction, (2) attribute-object disentanglement, and (3) feature alignment.

Visual Feature Extraction

As shown in Figure 2, we denote a given image with the attribute-object composition label (e.g. `ripe apple`) as the main image x^m, and randomly sample an image with the same attribute x^a (*i.e.*, `ripe orange`) and an image sharing the same object x^o (*i.e.*, `peeled apple`) to form a triplet image set. For convenience of expression, we simply use x^{img} (where $img \in \{m, a, o\}$) to collectively denote the images as they are processed using the same module.

Visual backbone. As mentioned before, since LLaVA v1.5 is used as our fundamental MLLM, we directly leverage its visual encoder (*i.e.*, ViT) and cross-modal connector (CMC) to extract visual features. Specifically, the image x^{img} is partitioned into n patch tokens, which are subsequently put into ViT along with the `[CLS]` token. Afterward, the output of patch tokens before the last layer of ViT is fed into the CMC module, as implemented in LLaVA v1.5. To align the dimensions of the patch tokens output by CMC and the `[CLS]` token produced by ViT, the patch tokens output by CMC is input into a linear layer. Consequently, we obtain one feature vector of `[CLS]` token $f_{cls}^{img} \in \mathbb{R}^d$ and a patch feature matrix of n patch tokens $F_{patch}^{img} \in \mathbb{R}^{n \times d}$.

Local features extraction. Intuitively, the composition (*e.g.*, `ripe apple`) only occupies a few parts of the image. Since each patch token corresponds to one local region of the image, to filter out background noise and focus on related regions, we deploy p feature adaptive aggregation (FAA) modules to derive p relevant local features of x^{img}, where each FAA module is formulated as follows:

$$\begin{cases} v = agg \otimes F_{patch}^{img} \\ agg = ReLU(Conv(F_{patch}^{img})) \end{cases} \quad (1)$$

where $Conv(\cdot)$ represents the 1×1 convolution layer, $agg \in \mathbb{R}^n$ is the weight vector, and the k-th element of agg is the weight for k-th patch feature. $\otimes$ represents matrix product, and $v \in \mathbb{R}^d$ is the local feature obtained by an FAA module. We vertically concatenate the local features produced by p FAA modules to obtain the local feature matrix $F_l^{img} \in \mathbb{R}^{p \times d}$.

Global features extraction. Normally, the `[CLS]` token output by ViT is regarded as containing various global information of the image, which highly entangles both attribute and object together [Hao *et al.*, 2023]. To disperse multi-granularity global information into different representations at the dimensional level, q learnable condition masks are applied to f_{cls}^{img} to obtain q different global representations. Each global representation is computed as:

$$u = f_{cls}^{img} \odot c \tag{2}$$

where $u \in \mathbb{R}^d$ denotes the global representation, $c \in \mathbb{R}^d$ refers to the learnable condition mask and $\odot$ is the element-wise multiplication. We vertically concatenate q global representations to derive the global feature matrix $F_g^{img} \in \mathbb{R}^{q \times d}$.

Features concatenation. Finally, F_l^{img} and F_g^{img} are vertically concatenated to form the visual features of x^{img}, *i.e.*, $F^{img} = [F_l^{img}, F_g^{img}] \in \mathbb{R}^{h \times d}$ (where $h = p + q$), which is used for the following attribute-object disentanglement.

Orthogonal regularization. Different features should capture distinct and complementary aspects of the image. Therefore, we introduce the orthogonal regularization, i.e.:

$$\mathcal{L}_{ortho} = \frac{1}{|img| \cdot |i|} \sum_{img \in \{m,a,o\}, i \in \{g,l\}} (||F_i^{img} F_i^{img^T} - I_i||_{Fro}) \tag{3}$$

where I_i denotes the identity matrix, and $||\cdot||_{Fro}$ refers to the Frobenius norm of the matrix.

Image embedder. Inspired by [Nagarajan and Grauman, 2018], for the input image x^{img}, we first use the average pooling $Avg(\cdot)$ on F_g^{img} and F_l^{img}, respectively, and horizontally concatenate them by $Cat(\cdot, \cdot)$ to aggregate both global and local features of x^{img}. Then the concatenated feature passes through a linear layer $Lin_{comp}(\cdot)$ to derive the final visual feature f_{comp}^{img} that represents the composition. This module is formulated as follows:

$$f_{comp}^{img} = Lin_{comp}(Cat(Avg(F_g^{img}), Avg(F_l^{img}))) \tag{4}$$

Attribute-Object Disentanglement

As mentioned before, one of the key challenges for CZSL is to disentangle attributes and objects from visual features. To overcome the challenge, we propose a novel weighted disentanglement strategy, as illustrated in Figure 2. For brevity, the image pair (x^m, x^a) from the triplet image set is taken as an example to elaborate on this strategy, while another image pair (x^m, x^o) is processed in the same manner.

Weights computation. The features of x^m and x^a (*i.e.*, F^m and F^a) are vertically concatenated and fed into two MLP modules to derive their respective weights of shared attribute features relative to each other. Subsequently, we utilize them to compute the weights of their own exclusive object features as follows:

$$\begin{cases} w_{attr}^{m2a} = \sigma(MLP_{m2a}([F^m, F^a])) \\ w_{obj}^{m2a} = \mathbf{1} - w_{attr}^{m2a} \\ w_{attr}^{a2m} = \sigma(MLP_{a2m}([F^m, F^a])) \\ w_{obj}^{a2m} = \mathbf{1} - w_{attr}^{a2m} \end{cases} \tag{5}$$

where $w_{attr}^{m2a}, w_{attr}^{a2m} \in \mathbb{R}^h$ represent the weights of the shared attribute features of x^m relative to x^a and x^a relative to x^m, respectively. w_{obj}^{m2a} and w_{obj}^{a2m} denote the weights of exclusive object features corresponding to x^m and x^a, respectively, which are derived by "$1 - shared\ weights$" paradigm as beyond the shared features of the image pair are the exclusive features of each image. Taking w_{attr}^{m2a} as an example, its k-th element refers to the shared attribute proportion of the k-th feature of x^m relative to x^a.

Disentangled features acquisition. We multiply the elements of each weight vector by the corresponding features and then calculate the average. The following takes the process of obtaining the shared attribute features of image x^m relative to x^a as an example:

$$f_{attr}^{m2a} = \frac{1}{h} \sum_{i=1}^{h} w_{attr\,i}^{m2a} \, {F^m}_{i,:} \tag{6}$$

where ${F^m}_{i,:}$ denotes the i-th row of F^m, *i.e.*, the i-th feature of x^m. $w_{attr\,i}^{m2a}$ refers to the i-th element of w_{attr}^{m2a}.

For the image pair of x^m and x^a, four parts are obtained: the shared attribute features of x^m relative to x^a and x^a relative to x^m, as well as two exclusive object features of the two images, respectively. These four features are marked as f_{pri}^e, where $e \in \{m2a, a2m\}$ and $pri \in \{attr, obj\}$. Then the shared attribute feature of x^a and x^m without relativity is obtained by an MLP layer, which is less dependent on the different objects of the two images. The process is as follows:

$$f_{attr}^{ma} = MLP_{ma}(Cat(f_{attr}^{m2a}, f_{attr}^{a2m})) \tag{7}$$

Similarly, we disentangle the attribute and object for the image pair (x^m, x^o) and obtain the same visual features as (x^m, x^a): f_{pri}^e, where $e \in \{m2o, o2m\}$ and $pri \in \{obj, attr\}$, and the feature without relativity f_{obj}^{mo}.

Feature Alignment

Inspired by GritLM [Muennighoff *et al.*, 2024] that leverages the last hidden states of LLMs as the representational embeddings, we consider the last hidden states of LLaVA v1.5 [Liu *et al.*, 2024a] as our MLLM embeddings for primitive words. Moreover, to tackle the problem that the ineffective overconfidence exhibited by the model in terms of the ground-truth attribute hinders it from generalizing to unseen compositions, we employ GPT-3.5 [OpenAI, 2023] to generate auxiliary attributes for compositions and perform label smoothing during attribute alignment. The auxiliary attributes and MLLM embeddings are obtained offline before training TRIDENT.

Auxiliary attributes generation by LLM. Since only textual attributes need to be generated, the LLM GPT-3.5 [OpenAI, 2023], instead of an MLLM, is leveraged to generate t auxiliary attributes for each composition. Specifically, the

following prompt is input to LLM: '*Please give me t adjectives that can describe the visual feature of a photo of a/an ... well.*', where the attribute-object composition (*e.g.*, `peeled apple`) is filled in '...'. Subsequently, the generated auxiliary attribute words form a set $\mathcal{A}_a$. Therefore, the set of all words $\mathcal{Y}$ is obtained, including attributes, objects and auxiliary attributes:

$$\mathcal{Y} = \mathcal{A} \cup \mathcal{O} \cup \mathcal{A}_a \tag{8}$$

MLLM embeddings acquisition. Each word $y \in \mathcal{Y}$ is fed into LLaVA v1.5 to get the last hidden states, *i.e.*, $LLaVA_{lhs}(\cdot)$. Specifically, y is tokenized into multiple subwords and passed through the MLLM; the final-layer output is averaged as the MLLM embedding of the word. Subsequently, it is passed through an MLP layer to obtain the embedding $E_{word}(\cdot)$ of the aligned dimension with visual features. And for a composed pair c of attribute a and object o, *i.e.*, $c = (a, o)$, we horizontally concatenate the MLLM embeddings for a and o and feed them into a linear layer $Lin_{co}(\cdot)$ to get the composed pair embedding $E_{co}(\cdot)$. The process is formulated as follows:

$$E_{word}(y) = MLP_{word}(LLaVA_{lhs}(y)) \tag{9}$$

$$E_{co}(c) = Lin_{co}(Cat(LLaVA_{lhs}(a), (LLaVA_{lhs}(o))) \tag{10}$$

Word expanding. Previous works compute cosine similarities of disentangled features and word embeddings and apply cross-entropy only within the respective domains of attributes or objects, which results in the disentangled attributes and objects still retaining the information of each other. To address the problem, we propose a novel word expanding strategy, which computes cosine similarities of visual features and the embeddings of all words, including attributes and objects, and treats all words except the ground-truth word as the negative labels in subsequent cross-entropy.

Alignment by cross-entropy. Similar to [Mancini *et al.*, 2021], we use cross-entropy to align the visual features and word embeddings. Assume that f is the visual feature and $E_{word}(wd)$ is the word embedding for the word $wd \in \mathcal{Y}$. The classifier logit from f to $E_{word}(wd)$ is defined as follows:

$$CE(f, wd) = \frac{e^{\delta \cdot \cos(f, E_{word}(wd))}}{\sum_{wd' \in \mathcal{Y}} e^{\delta \cdot \cos(f, E_{word}(wd'))}} \tag{11}$$

where δ is the temperature variable, and $cos(\cdot, \cdot)$ denotes the cosine similarity function. Thus cross-entropy with/without label smoothing can be uniformly formulated as follows:

$$H(f, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} -z \log(CE(f, y))$$

$$\text{with } z = \begin{cases} 1-\alpha, & \text{if } y \text{ is ground truth label} \\ \alpha/t, & \text{if } y \text{ is auxiliary label} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where α denotes the smoothing factor, t refers to the number of auxiliary labels, and $z \in [0, 1]$ represents the target value of one-hot label or smoothing label. For cross-entropy without label smoothing, *i.e.*, with one-hot label H_{oh}, α is set to 0. The cross-entropy with label smoothing is denoted as H_{ls}.

Dataset	*Primitive*		*Train*		*Validation*			*Test*		
	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{O}\rvert$	$\lvert\mathcal{C}_s\rvert$	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{C}_s\rvert$	$\lvert\mathcal{C}_u\rvert$	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{C}_s\rvert$	$\lvert\mathcal{C}_u\rvert$	$\lvert\mathcal{X}\rvert$
MIT-States	115	245	1262	30k	300	300	10k	400	400	13k
C-GQA	413	674	5592	27k	1252	1040	7k	888	923	5k
VAW-CZSL	440	541	1252	72k	2121	2322	10k	2449	2470	11k

Table 1: Summary statistics of the datasets used in our experiments.

For the disentangled attribute features of one image relative to each other, since a single object exhibits multiple attributes, we exploit attribute smoothing with auxiliary attributes to undermine the overconfidence in the ground-truth attribute and learn more related attributes. For the shared attribute features without relativity, the one-hot label is used to learn a pure attribute concept that is less conditioned on objects. The loss for disentangled attributes is as follows:

$$\mathcal{L}_{attr} = \frac{1}{|e|+1}\Big(\sum_{e \in \{m2a, a2m, m2o, o2m\}} H_{ls}(f^e_{attr}, \mathcal{Y}) + H_{oh}(f^{ma}_{attr}, \mathcal{Y})\Big) \tag{13}$$

Concerning the disentangled object features, we use cross-entropy with one-hot label to learn the object concept and the loss is as follows:

$$\mathcal{L}_{obj} = \frac{1}{|e|+1}\Big(\sum_{e \in \{m2a, a2m, m2o, o2m\}} H_{oh}(f^e_{obj}, \mathcal{Y}) + H_{oh}(f^{mo}_{obj}, \mathcal{Y})\Big) \tag{14}$$

With respect to the visual feature of the composition from image embedder, we calculate the cosine similarities between the visual embedding and the composed pair embeddings of seen composition labels. The cross-entropy loss for the composition is as follows:

$$\mathcal{L}_{comp} = \frac{1}{|img|} \sum_{img \in \{m,a,o\}} H_{oh}(f^{img}_{comp}, \mathcal{C}_s) \tag{15}$$

3.3 Training and Inference

During the training phase, the overall loss is as follows:

$$\mathcal{L} = \gamma_{ortho}\mathcal{L}_{ortho} + \gamma_{comp}\mathcal{L}_{comp} + \gamma_{pri}(\mathcal{L}_{attr} + \mathcal{L}_{obj})/2 \tag{16}$$

where γ_{ortho}, γ_{comp}, and γ_{pri} are weighting factors to balance the influence of different losses.

For inference, given an image from the test set, the cosine similarities of its visual feature obtained by image embedder and the composed pair embeddings of all candidate compositions are computed. The composition with the highest similarity is predicted by the model. Note that although the disentanglement branch is not used for inference, it still influences the formation of the composition feature space through the visual feature extraction module described in Section 3.2.

4 Experiment

4.1 Experiment Setup

Datasets. We evaluate our model on three challenging CZSL datasets: MIT-states [Isola *et al.*, 2015], C-GQA [Naeem *et al.*, 2021], and VAW-CZSL [Saini *et al.*, 2022]. The common data splits are presented in Table 1.

Metrics. Our method is evaluated on seen and unseen compositions separately. Following the common generalized CZSL setting [Purushwalkam *et al.*, 2019], a calibration bias

Method	MIT-States				C-GQA				VAW-CZSL			
	AUC	*HM*	*Seen*	*Unseen*	*AUC*	*HM*	*Seen*	*Unseen*	*AUC*	*HM*	*Seen*	*Unseen*
SymNet [Li *et al.*, 2020]	3.2	13.7	22.7	20.1	1.9	10.8	20.3	11.8	2.8	13.5	20.2	18.0
CompCos [Mancini *et al.*, 2021]	12.3	28.2	39.0	39.5	5.0	17.7	32.8	19.1	6.5	20.8	30.5	27.4
Co-CGE [Mancini *et al.*, 2022]	10.3	25.1	41.0	33.1	4.2	15.2	32.9	17.0	6.2	19.7	31.0	26.1
SCEN [Li *et al.*, 2022]	9.8	24.6	35.1	36.5	3.8	15.3	31.5	15.7	5.7	19.2	29.9	24.5
OADis [Saini *et al.*, 2022]	13.1	29.0	42.3	27.3	2.3	12.1	23.3	12.8	4.1	16.2	26.0	20.7
INV [Zhang *et al.*, 2022]	11.5	26.6	28.5	25.0	1.4	7.9	28.6	6.8	2.0	11.1	21.1	11.9
CANet [Wang *et al.*, 2023]	13.6	29.8	**46.4**	39.9	5.7	18.9	34.8	20.5	6.7	21.0	31.2	27.4
ProCC [Huo *et al.*, 2024]	9.5	28.1	43.1	39.1	3.5	15.1	32.4	15.8	3.6	18.9	26.9	25.5
CLIP [Nayak *et al.*, 2023]	11.0	26.1	30.2	46.0	1.4	8.6	7.5	25.0	-	-	-	-
CoOp [Nayak *et al.*, 2023]	13.5	29.8	34.4	**47.6**	4.4	17.1	20.5	**26.8**	-	-	-	-
TRIDENT (Ours)	**14.2**	**30.9**	44.5	40.0	**8.0**	**22.6**	**39.5**	24.1	**8.3**	**23.4**	**33.3**	**31.1**

Table 2: Comparison with the state-of-the-art results on MIT-States, C-GQA and VAW-CZSL. The four indicators are explained in Metrics. We measure top-1 AUC on MIT-States and C-GQA, and top-3 AUC on VAW-CZSL. The best results are displayed in **boldface**, and the second best results are underlined.

Method	*AUC*	*HM*	*Seen*	*Unseen*
w/o patch_features	12.9	28.9	42.6	38.0
w/o [CLS]_feature	13.4	30.1	44.3	39.7
w/o FAAs	13.9	30.4	44.4	39.7
w/o condition_masks	14.0	30.5	44.2	39.8
w/o word_expanding	14.0	30.1	44.7	39.8
w/o attribute_smoothing	13.9	30.5	**44.9**	39.5
w/o $\mathcal{L}_{attr} + \mathcal{L}_{obj}$	13.2	30.1	43.8	38.9
w/o $\mathcal{L}_{ortho}$	14.1	30.7	44.6	39.7
TRIDENT	**14.2**	**30.9**	44.5	**40.0**

Table 3: Ablation study results on MIT-States. *w/o* {*certain_part*} denotes this part is ablated.

Method	*Varient*	*AUC*	*HM*
SCEN [Li *et al.*, 2022]	*ft+w2v*	8.2	22.8
	LLaVA$_{lhs}$	**10.3**	**25.1**
CANet [Wang *et al.*, 2023]	*ft+w2v*	12.3	**28.4**
	LLaVA$_{lhs}$	**12.5**	28.3
TRIDENT	*ft+w2v*	14.0	29.9
	LLaVA$_{lhs}$	**14.2**	**30.9**

Table 4: Impact of word embedding on MIT-States. $ft+w2v$ means the sum of Word2Vec and Fasttext. $LLAVA_{lhs}$ represents the last hidden states of LLAVA v1.5.

trades off between the accuracies of seen and unseen compositions. We calculate the Area Under the Curve (AUC) using seen and unseen accuracies at different biases. The best $Seen$ and $Unseen$ accuracies of the curve are also reported. In addition, we calculate the Harmonic Mean of seen and unseen accuracies at different biases and report the best one (HM).

Implementation details. We use the visual encoder of LLaVA v1.5, ViT-Large-Patch14-336px as our frozen visual backbone. **TRIDENT** and all baseline models are trained with the batch size of 128 for 50 epochs under the PyTorch framework [Paszke *et al.*, 2019]. The number of global features is set to 6, 2, and 4 for the three datasets, respectively, and the number of local features is twice that of the global features. The label smoothing factor is set to 0.09, 0.03, and 0.03 for the three datasets, respectively. The number of auxiliary attributes generated for each composition is set to 3. We train **TRIDENT** by Adam optimizer with the weight decay of 5e-5, learning rates of 1.5e-6 for word embedding and 2e-4 for other modules. We decay the learning rate by a factor of 0.1 at epoch 30 and 40. The temperature variable of cosine similarity δ is set to 0.05. For weighting coefficients γ_{ortho}, γ_{comp}, and γ_{pri}, we set them to 0.1, 1, 0.25, respectively.

Baselines. We compare our **TRIDENT** with recent and prominent approaches in CZSL: SymNet [Li *et al.*, 2020], CompCos [Mancini *et al.*, 2021], Co-CGE [Mancini *et al.*, 2022], SCEN [Li *et al.*, 2022], OADis [Saini *et al.*, 2022], INV [Zhang *et al.*, 2022], CANet [Wang *et al.*, 2023], and ProCC [Huo *et al.*, 2024]. We replace their visual backbone with ViT-Large-Patch14-336px and retrain all models with the same number of epochs for the sake of fairness. In addition, although comparing **TRIDENT** with CLIP-based methods, which rely on the dual-tower architecture, is very unfair due to inadvertent exposure to unseen compositions, we still choose CLIP [Radford *et al.*, 2021] and CoOp [Zhou *et al.*, 2022] as baselines for their strong zero-shot abilities.

4.2 Results and Discussion

In this section, we compare **TRIDENT** with state-of-the-art methods. As shown in Table 2, our model surpasses other models by a substantial margin in general. **TRIDENT** boosts AUC from 13.6%, 5.7%, and 6.7% of the previous state-of-the-art method CANet to the new state-of-the-art performance of 14.2%, 8.0%, and 8.3% with 0.6%, 2.3%, and 1.6% improvement on three datasets, respectively. In addition, **TRIDENT** achieves 30.9%, 22.6%, 23.4% on the metrics of HM, providing 1.1%, 3.7%, and 2.4% improvement on CANet. For MIT-States, our model achieves competitive performance, despite considerable label noise [Atzmon *et al.*, 2020]. The largest improvement is observed on the $Unseen$ metric, indicating that attribute smoothing helps enhance the generalization ability of the model. We also observe that **TRIDENT** performs significantly better than CANet regarding all metrics on two more challenging and low-noise datasets C-GQA and VAW-CZSL, indicating the efficacy of our approach. These improvements arise from the utilization of MLLM embeddings and attribute smooth-

(a) image-to-text retrieval. (b) text-to-image retrieval (successful cases). (c) text-to-image retrieval (failure cases).

Figure 3: Qualitative analysis. (a) Top-5 image-to-text retrieval. The first two rows display successful cases, while the last row presents failure cases. For each image, the top row shows the ground-truth, followed by five rows of top-5 predictions. (b) Successful cases and (c) failure cases of top-5 text-to-image retrieval. In all cases, the successful and failure results are tagged in green and red, respectively.

ing, which enhance attribute-object disentanglement and consequently facilitate the recognition of unseen compositions while maintaining performance on seen compositions.

4.3 Ablation Study

We ablate the components of **TRIDENT** on MIT-States to evaluate their contributions. From the ablation results in Table 3, we gain the following observations.

1) **TRIDENT** outperforms the models without using patch and `[CLS]` features, indicating that both patch and `[CLS]` features are crucial, with patch features contributing more.

2) Both *w/o FAAs* and *w/o condition_masks* models perform worse than **TRIDENT**, which validates the importance of filtering out the background noise and extracting the multi-granularity features, respectively.

3) **TRIDENT** surpasses *w/o word_expanding* model and *w/o attribute_smoothing* model on the $Unseen$ metric, yet falls short of them on the $Seen$ metric. The difference between our model and the *w/o word_expanding* model stems from its more thorough disentanglement, which enhances the recognition of unseen compositions while weakening the identification of seen ones. The disparity between our model and the *w/o attribute_smoothing* model arises from attribute smoothing, which diminishes the overconfidence of the model in seen compositions, facilitating its generalization to unseen ones. However, the improvement of our model over these two models on AUC and HM indicates the effectiveness of the word expanding and label smoothing strategy.

4) **TRIDENT** outperforms *w/o* $\mathcal{L}_{attr} + \mathcal{L}_{obj}$ model on all metrics, confirming that the attribute-object disentanglement module is highly advantageous for generalization from seen compositions to unseen compositions.

5) w/o $\mathcal{L}_{ortho}$ model is inferior to **TRIDENT**, which suggests that the designed orthogonal regularization is helpful to guarantee that different features extract different visual information.

Impact of word embeddings. Our work leverages the last hidden states of LLaVA v1.5 as word embeddings, while Word2Vec [Mikolov, 2013] and FastText [Bojanowski *et al.*, 2017] are the common word embeddings used for MIT-States in previous works. In Table 4, based on three models: SCEN [Li *et al.*, 2022], CANet [Wang *et al.*, 2023] and **TRIDENT**, we compare the performance of using the last hidden states of LLaVA v1.5 ($LLaVA_{lhs}$) and the sum of Word2Vec and FastText (*ft+w2v*), respectively. The results indicate that the last hidden states of MLLM capture more complex multimodal semantic information than ordinary word embeddings.

4.4 Qualitative Analysis

In this section, we use **TRIDENT** to conduct both image-to-text retrieval and text-to-image retrieval experiments on the three datasets. We first consider image-to-text retrieval, shown in Figure 3a. For successful cases, such as the image labeled `burnt house`, we notice that the top four predictions can describe logs burning on a fire in the image. In terms of the image labeled `green leaf`, another successful case, all predicted attributes can describe the `leaf`, which is due to attribute smoothing learning more attributes for an object. For the failure cases, such as the image labeled `colorful pillow`, the prediction of `orange pillow` can also describe the image.

We then consider text-to-image retrieval. Successful cases are shown in Figure 3b, while failure cases are shown in Figure 3c. We observe that all retrieved images of `peeled orange` are correct. However, the retrieved images of `green grapes` are all wrong. This is due to the fact that the training images of `green grapes` are almost entirely filled with a single grape, which makes it difficult for the model to capture the contour features of a bunch of `green grapes`.

5 Conclusion

In this work, we propose a novel framework termed **TRIDENT** to address the challenging CZSL task. First, we leverage feature adaptive aggregation modules to mitigate the impact of background, and utilize learnable condition masks to capture multi-granularity features for attribute-object disentanglement. In addition, we exploit the last hidden states of MLLM to replace ordinary word embeddings, as they can capture complex multimodal semantic information. Moreover, we leverage LLM to generate auxiliary attributes and perform attribute smoothing to diminish overconfidence of the model in seen compositions, which enables it to generalize to unseen compositions more effectively. Extensive experiments conducted on three challenging datasets demonstrate the effectiveness of our method.

Acknowledgments

Our work was supported by the Beijing Natural Science Foundation (No. 4242046).

References

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736, 2022.

[Atzmon *et al.*, 2020] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In *Advances in Neural Information Processing Systems*, volume 33, pages 1462–1473, 2020.

[Bojanowski *et al.*, 2017] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Hao *et al.*, 2023] Shaozhe Hao, Kai Han, and Kwan-Yee K. Wong. Learning attention as disentangler for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15315–15324, 2023.

[Huang *et al.*, 2024] Siteng Huang, Biao Gong, Yutong Feng, Min Zhang, Yiliang Lv, and Donglin Wang. Troika: Multi-path cross-modal traction for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24005–24014, 2024.

[Huo *et al.*, 2024] Fushuo Huo, Wenchao Xu, Song Guo, Jingcai Guo, Haozhao Wang, Ziming Liu, and Xiaocheng Lu. Procc: Progressive cross-primitive compatibility for open-world compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12689–12697, 2024.

[Isola *et al.*, 2015] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1383–1391, 2015.

[Li *et al.*, 2020] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11313––11322, 2020.

[Li *et al.*, 2022] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. Siamese contrastive embedding network for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9335, 2022.

[Li *et al.*, 2024] Xiangyu Li, Xu Yang, Xi Wang, and Cheng Deng. Agree to disagree: Exploring partial semantic consistency against visual deviation for compositional zero-shot learning. *IEEE Transactions on Cognitive and Developmental Systems*, 16(4):1433–1444, 2024.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[Liu *et al.*, 2024b] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[Lu *et al.*, 2023] Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23560–23569, 2023.

[Mancini *et al.*, 2021] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5222–5230, 2021.

[Mancini *et al.*, 2022] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1545–1560, 2022.

[Mikolov, 2013] Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Misra *et al.*, 2017] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1160–1169, 2017.

[Muennighoff *et al.*, 2024] Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*, 2024.

[Muennighoff, 2022] Niklas Muennighoff. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*, 2022.

[Naeem *et al.*, 2021] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962, 2021.

[Nagarajan and Grauman, 2018] Tushar Nagarajan and Kristen Grauman. Attributes as operators: Factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision*, pages 169–185, 2018.

[Nan *et al.*, 2019] Zhixiong Nan, Yang Liu, Nanning Zheng, and Song-Chun Zhu. Recognizing unseen attribute-object pair with generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8811–8818, 2019.

[Nayak *et al.*, 2023] Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. Learning to compose soft prompts for compositional zero-shot learning. In *Proceedings of the International Conference on Learning Representations*, 2023.

[OpenAI, 2023] OpenAI. Gpt-3.5-turbo api. https://platform.openai.com/docs/models/gpt-3-5, 2023. Access: 2024-7-10.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[Purushwalkam *et al.*, 2019] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3592–3601, 2019.

[Radford *et al.*, 2019] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021.

[Ruis *et al.*, 2021] Frank Ruis, Gertjan Burghouts, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. In *Advances in Neural Information Processing Systems*, volume 34, pages 10641–10653, 2021.

[Saini *et al.*, 2022] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667, 2022.

[Sarzynska-Wawer *et al.*, 2021] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135, 2021.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Wang and Kuo, 2020] Bin Wang and C.-C. Jay Kuo. Sbert-wk: A sentence embedding method by dissecting bert-based word models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2146–2157, 2020.

[Wang *et al.*, 2023] Qingsheng Wang, Lingqiao Liu, Chenchen Jing, Hao Chen, Guoqiang Liang, Peng Wang, and Chunhua Shen. Learning conditional attributes for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2023.

[Xu *et al.*, 2024] Shuo Xu, Sai Wang, Xinyue Hu, Yutian Lin, Bo Du, and Yu Wu. Mac: A benchmark for multiple attributes compositional zero-shot learning. *arXiv preprint arXiv:2406.12757*, 2024.

[Zhang *et al.*, 2022] Tian Zhang, Kongming Liang, Ruoyi Du, Xian Sun, Zhanyu Ma, and Jun Guo. Learning invariant visual representations for compositional zero-shot learning. In *Proceedings of the European Conference on Computer Vision*, pages 339–355, 2022.

[Zhang *et al.*, 2024] Yang Zhang, Songhe Feng, and Jiazheng Yuan. Continual compositional zero-shot learning. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1724–1732, 2024.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.