Contextual Feature Expansion with Superordinate Concept for Compositional Zero-Shot Learning

Soohyeong Kim and Yong Suk Choi *

Posted Date: 7 August 2025

doi: 10.20944/preprints202508.0451.v1

Keywords: compositional zero-shot learning (CZSL); superordinate concept representation; fuzzy logic; spectral clustering

 doi:10.20944/preprints202508.0451.v1

Article

Contextual Feature Expansion with Superordinate Concept for Compositional Zero-Shot Learning

Soohyeong Kim [1] **and Yong Suk Choi** *

Department of Artificial Intelligence, Hanyang University, Seoul 04763, Republic of Korea
* Correspondence: cys@hanyang.ac.kr

Abstract

Compositional Zero-Shot Learning (CZSL) seeks to enable machines to recognize objects and attributes (*i.e.*, *primitives*), learn their associations, and generalize to novel compositions, enabling systems to exhibit a human-like ability to infer and generalize. Existing approaches, multi-label and multi-class classification, face inherent t rade-offs: the former suffers from biases against unrelated compositions, while the latter struggles with exponentially growing search spaces as the number of objects and attributes increases. To overcome these limitations and address the exponential complexity in CZSL, we introduce **C**oncept-**o**riented **F**eature **AD**justment (CoFAD), a novel method that extracts superordinate conceptual features based on primitive relationships and expands label feature boundaries. By incorporating spectral clustering and membership function in fuzzy logic, CoFAD achieves state-of-the-art performance while using 2×–4× less GPU memory and reducing training time by up to 50× on large-scale dataset.

Keywords: compositional zero-shot learning (CZSL); superordinate concept representation; fuzzy logic; spectral clustering

1. Introduction

In nature, all objects or entities are associated with their respective superordinate concepts, and the attributes applied to these objects are strongly correlated with their corresponding superordinate categories [1,2]. For instance, water and juice, both belonging to the superordinate concept of liquid, can exhibit the attribute "spilled". In contrast, the attribute "bright" might apply to both "lamp" (an object) and "butterfly" (an animal), yet these belong to vastly different superordinate categories. Learning the relationships between concepts forms the foundation of human cognition and represents an essential challenge for machine learning. Similarly, Compositional Zero-Shot Learning (CZSL) [3–7] aims to replicate this nuanced understanding in machines by enabling them to associate objects, attributes, and their superordinate categories. Recent advancements have transitioned CZSL from the traditional closed-world setting [3,4], which focuses on predicting existing combinations, to an open-world setting [8–10] that includes impossible combinations (e.g., rusty dog) in the search space. This transition accounts for the unpredictability and variability of real-world data, making CZSL more robust and applicable to practical scenarios.

Table 1. Data splits for the three benchmark datasets used in this work. $|A|$ and $|O|$ represent the number of attributes and objects, repectively. Y_S represents the number of seen compositions, and $|Y_U|$ represents the number of unseen compositions. X denotes the number of image samples used in each split.

Dataset	$\|A\|$	$\|O\|$	$\|A\|\times\|O\|$	Training		Validation			Test		
				Y_S	X	Y_S	Y_U	X	Y_S	Y_U	X
UT-Zappos	16	12	192	83	23k	15	15	3k	18	18	3k
MIT-States	115	245	28,175	1262	30k	300	300	10k	400	400	13k
CGQA	413	674	278,362	5592	27k	1040	1252	7k	888	923	5k

 doi:10.20944/preprints202508.0451.v1

Most studies on the CZSL task employ multi-label [5,8,11], multi-class [6,12,13], or multi-path [14,15] classification approaches. Multi-label classification predicts attributes and objects for a given input image. This disentanglement of compositions reduces computational cost to the number of primitives and eliminates the need to explicitly predict unseen compositions during inference. However, when loosely related superordinate concepts share similar attributes, this approach tends to be biased by the training data, leading to poor recognition of features in unseen compositions. The multi-class classification approach improves performance by incorporating all possible compositions into the model's search space, allowing it to learn relationships between compositions. However, as the number of attributes ($|A|$) and objects ($|O|$) increases, the search space grows exponentially to $|A| \times |O|$, leading to a substantial increase in computational cost. Table 1 clearly quantifies the increase in the search space. This issue also arises in the multi-path classification approach, in which the model predicts primitives and compositions simultaneously. The trade-off between performance and efficiency is a significant challenge in imitating human-like intelligence in real-world scenarios.

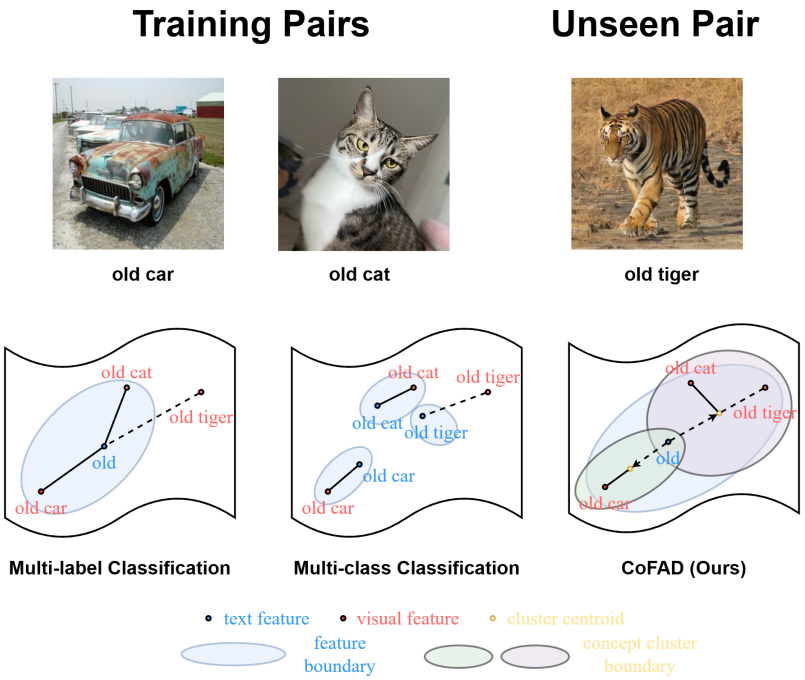

Figure 1. Comparison of label feature spaces between conventional multi-label and multi-class strategies and the proposed CoFAD method

To overcome these limitations and enable more efficient recognition of unseen compositions, we propose a method called **C**oncept-**o**riented **F**eature **AD**justment (CoFAD), which expands a primitive's feature space to include unseen compositions using its associated superordinate concepts. CoFAD extracts the superordinate concepts of primitives based on the connection of compositions. These superordinate concepts represent shared features of associated primitives; for instance, "cat," "tiger," and "old" share the concept of creature or animal, while "car" and "truck" are associated with the superordinate concept of vehicle. CoFAD expands the feature space of primitives by assigning membership degrees to each superordinate concept, thereby allowing primitives to implicitly include other primitive features and correcting biases present in seen compositions. A significant advantage of this approach is the reduction of the model's search space into $|A| + |O|$, which dramatically decreases computational cost while simultaneously improving performance.

We conduct evaluations on three popular benchmark datasets: MIT-States [16], UT-Zappos [17], and C-GQA [7]. Our proposed CoFAD model demonstrates superior performance, surpassing state-of-the-art (SOTA) results in open-world scenarios. By leveraging diverse learning strategies, CoFAD effectively reduces the search space, achieving competitive performance even in closed-world settings. Notably, CoFAD offers exceptional computational efficiency compared to existing models, requiring

 doi:10.20944/preprints202508.0451.v1

2×–4× less GPU memory and achieving training time reductions ranging from 3× to 50×. The primary contributions of this work are as follows:

- Novel Model Architecture:We introduce the CoFAD model, which extends contextual feature boundaries through concept-oriented learning.
- **State-of-the-Art Performance**: Experimental results on benchmark datasets establish CoFAD as SOTA in open-world scenarios, while maintaining strong performance in closed-world settings.
- **Enhanced Computational Efficiency**: CoFAD achieves remarkable efficiency, utilizing significantly less GPU memory and reducing training times by up to 50× compared to SOTA models.

2. Related Work

Compositional Zero-Shot Learning (CZSL) seeks to identify unseen attribute-object combinations during testing and has evolved through three main approaches. The first approach, multi-label classification [6,12,13,18,19], trains separate classifiers to independently predict attributes and objects from image features, enabling the disentanglement of their representations. The second approach, multi-class classification [3,7,9], learns joint attribute-object representations for both seen and unseen combinations using transformation models such as multi-layer perceptrons (MLPs) [3,9] or graph convolutional networks (GCNs) [7]. The third and most recent approach, multi-path learning [14,15], learns compositions and primitives simultaneously, achieving significant improvements in performance. Recently, these approaches have increasingly incorporated connections between visual features and compositional representations, leveraging knowledge from pre-trained Vision-Language Models (VLMs). For instance, CSP [20] adapts the CLIP model [21] by replacing class-specific textual prompts with trainable attribute and object tokens. Similarly, Troika [14] introduces a Multi-Path Cross-Modal Traction module to generate prompts closely aligned with visual content. Building on these ideas, the CDS-CZSL [15] adopts a multi-path approach to capture the diversity and specificity of primitives in context.

Despite the progress achieved by these approaches, challenges persist, including limited generalization to unseen data and significant computational overhead, highlighting the need for innovative solutions. In this study, we introduce an efficient novel multi-label classification framework based on VLMs.

3. Preliminaries

3.1. Promblem Formulation

Compositional Zero-Shot Learning (CZSL) is concerned with modeling images as compositions of primitives, specifically attributes (e.g., Old) and objects (e.g., Car). The composition space in CZSL is defined as the Cartesian product of all possible attributes and objects, $Y = A \times O$, capturing all attribute-object combinations. This space is divided into two disjoint sets: seen compositions (Y_S) and unseen compositions (Y_U), such that $Y_S \cap Y_U = \emptyset$ and $Y_S \cup Y_U = Y$. During training, the model learns from a dataset $S = \{(x, y) | x \in X_S, y \in Y_S\}$, where each image x is labeled with a composition $y = (a, o)$. At test time, the model is required to predict labels for both seen and unseen compositions within the test label space $Y_{test} = Y_S \cup Y_U$. Depending on the evaluation setting, the task can be restricted to a predefined subset of unseen compositions in closed-world evaluation (CW-CZSL) or extended to all possible compositions in open-world evaluation (OW-CZSL). The goal of CZSL is to learn a model $f : X \rightarrow Y_{test}$ capable of recognizing images from novel attribute-object compositions.

3.2. Construct Prompt and Backbone

To leverage pre-trained knowledge from CLIP, following the approach in prior work [14], we construct prompts for each primitive using pre-trained embeddings from CLIP [21]. A new primitive vocabulary $V = [V_A, V_O] \in \mathbb{R}^{(|A|+|O|)\times dim}$ is created for all attributes and objects, where dim_t represents the embedding dimension of each token. To generate prompts for each primitive, we append a

 doi:10.20944/preprints202508.0451.v1

tokenized prefix, "a photo of", to the primitives. The prompt for each primitive is defined as $P_i = [p_1, \ldots, p_m, v_i]$, where $\{p_1, \ldots p_m\}$ are the prefix tokens, and all tokens are fully trainable. The text feature of a primitive is obtained by feeding the prompt P_i into the CLIP-based text encoder E^t, formulated as:

$$x_i^t = E^t(P_i). \tag{1}$$

Following prior works [14,15], we utilize an Adapter [22] for the visual encoder E^v, enabling adaptation without updating its parameters. The composition visual feature extracted from the encoder E^v is disentangled into two primitives via disentangler layers D_a and D_o, formulated as:

$$x_a^v = D_a(E^v(x)), x_o^v = D_o(E^v(x)). \tag{2}$$

4. Methodology

Approach. As discussed in the Introduction section, recent methodologies in CZSL often suffer from two critical limitations: models either fail to learn the relationships between primitives due to biases in the training set, or they incur substantial computational costs to handle the large search space of $|A| \times |O|$. We hypothesize that this trade-off between performance and computational cost arises from the design of current models, which focus on learning direct one-to-one associations between attribute-object pairs. To address this issue, we propose a novel method that enables the model to learn the relationships between a subordinate concept, primitive, and its associated superordinate concepts. Specifically, an attribute such as "old" manifests differently depending on the superordinate concept. For example, within animals, "old" may be characterized by increased body mass or fur growth, whereas in vehicles, it is reflected in features such as rust or outdated designs. However, subordinate concepts under the same superordinate category, such as truck and car within vehicles, exhibit similar characteristics for the "old" attribute. Based on this principle, we introduce a novel method **CoFAD**, **C**oncept-**o**riented **F**eature **AD**justment, that combines spectral clustering with the membership function in fuzzy logic. This approach enables the model to effectively learn the associations between primitives and the feature representations of their superordinate concepts. The overall framework of our proposed method is illustrated in Figure 2.

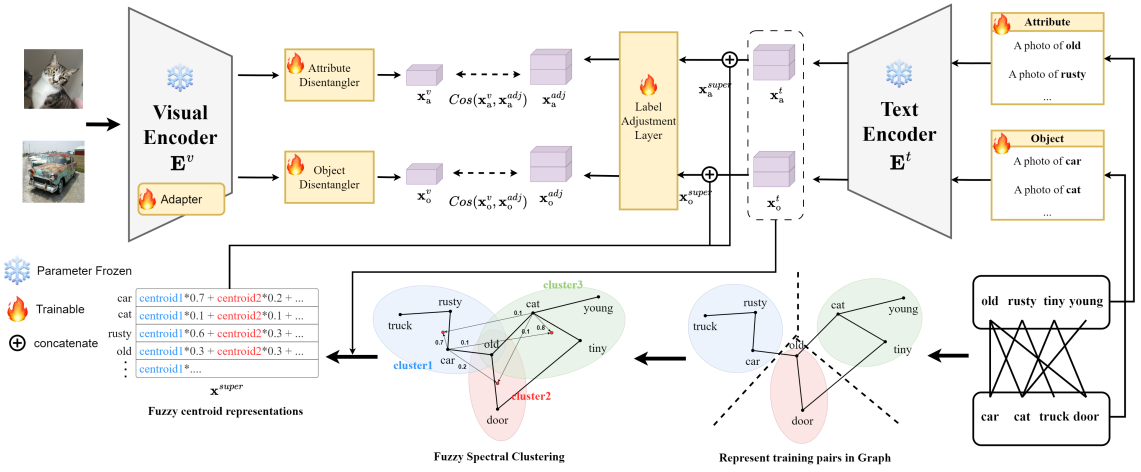

Figure 2. The overall flow of the proposed CoFAD method. Both the visual and text encoders are based on a CLIP-based model. Training pairs are represented as connections between labels, forming a graph structure. Fuzzy spectral clustering is then applied to derive superordinate concept features (Section 4.1). Labels are adjusted to align with superordinate concept features based on their memberships to these concepts, thereby expanding the feature space (Section 4.3). Based on cosine similarity, the label features are trained on potential unseen pair features derived from their respective superordinate concepts (Section 4.4).

 doi:10.20944/preprints202508.0451.v1

4.1. Fuzzy Spectral Clustering

Learning the relationships between primitives and their associated superordinate concepts requires identifying feature vectors that represent these superordinate concepts. In CZSL, a common approach involves minimizing the cosine similarity score between the visually encoded features and the word embeddings of attribute-object compositions. However, the pre-trained word embeddings used in this process (e.g., Word2Vec, FastText, or the embedding layer of the backbone network) often contain noise unrelated to the features present in the dataset, which can hinder accurate learning. To eliminate this noise and focus solely on the relationships between the given primitives, we adopt spectral clustering, which allows clustering based on the connectivity of primitives. Formally, given a set of attributes $A = \{a_0, a_1, \ldots, a_n\}$ and a set of objects $O = \{o_0, o_1, \ldots, o_m\}$, we can define a combined set $U = A \cup O$, where $U = \{u_0, ..., u_n, u_{n+1}, ..., u_{n+m}\}$. The adjacency matrix M for the compositions in the training set is defined as follows:

$$M_{i,j} = \begin{cases} 1 & \text{if } (u_i, u_j) \in Y_S, \\ 0 & \text{else.} \end{cases} \tag{3}$$

To represent the graph structure, the Laplacian matrix L is computed using the degree matrix D, where the diagonal entries are defined as $D[i,i] = \sum_j M[i,j]$. The normalized Laplacian matrix is then calculated as:

$$L = I - D^{-\frac{1}{2}} M D^{-\frac{1}{2}}, \tag{4}$$

where I denotes the identity matrix. Next, eigenvalue decomposition is performed on L:

$$Lv_q = \lambda_q v_q, \tag{5}$$

where λ_q represents the q-th eigenvalue and v_q is the corresponding eigenvector. The eigenvectors corresponding to the smallest η eigenvalues are selected to form the matrix V, where η is sames as the number of clusters determined for each dataset. The K-means clustering algorithm is then applied to the eigenvector matrix V.

The clusters and centroids obtained solely from the connectivity information among primitives represent the superordinate concepts and their respective representative values. Since a primitive can belong to multiple superordinate concepts rather than being restricted to a single one, the degree of membership to each superordinate concept is calculated based on Euclidean distance using the membership function in fuzzy logic. The Euclidean distance $d_{i,k}$ between label point i and the k-th cluster center is calculated as follows:

$$d_{i,k} = ||V[i] - \mu_k||, \tag{6}$$

where μ_k is the center of the k-th cluster. Finally, the fuzzy membership $u_{i,k}$ of each label eigenvector is determined based on the distances $d_{i,k}$:

$$u_{i,k} = \frac{1}{d_{i,k}^{\frac{2}{f-1}}} \Bigg/ \sum_{j=1} \frac{1}{d_{i,j}^{\frac{2}{f-1}}}, \tag{7}$$

where f is the fuzziness parameter. This fuzzy membership quantifies the degree of association of each label with a given cluster. In other words, fuzzy membership represents the degree to which each primitive belongs to the η superordinate concepts. A highly skewed membership indicates a specialized primitive with fewer associated compositions, while a more uniform membership suggests a common concept capable of exhibiting diverse characteristics.

 doi:10.20944/preprints202508.0451.v1

4.2. Feasibility Score

Spectral clustering effectively excludes noise present in word features, allowing the model to focus solely on the given label settings. However, it still relies exclusively on training set compositions, resulting in feasible unseen compositions being assigned a value of 0 in the adjacency matrix M. We believe that incorporating compositional feasibility for unseen connections into the computation of adjacency matrix M can improve the accuracy of spectral clustering. To achieve this, we adopt the compositional feasibility estimation method from previous work [8], which is based on the conjecture that similar objects share similar attributes, whereas dissimilar objects do not. For an unseen composition (a_{uc}, o_{uc}), the attribute a_{uc} and object o_{uc} are paired with a set of objects $O_{sc} = \{o_0, o_1, \ldots, o_i\}$ and attributes $A_{sc} = \{a_0, a_1, \ldots, a_j\}$, respectively, from the training set Y_S. The feasibility scores of the unseen composition with respect to the object o_{uc} and attribute a_{uc} are defined as follows:

$$\begin{aligned} \rho_{obj}(a_{uc}, o_{uc}) &= \max_{o \in O_{sc}} cos(x^t_{o_{uc}}, x^t_o), \\ \rho_{attr}(a_{uc}, o_{uc}) &= \max_{a \in A_{sc}} cos(x^t_{a_{uc}}, x^t_a), \end{aligned} \tag{8}$$

where x^t_o represents the text feature of o encoded by a pre-trained text encoder E^t and $cos(\cdot)$ denotes the cosine similarity function. The mixed feasibility score for the adjacency matrix is then defined as:

$$\rho_{uc} = max(\frac{\rho_{obj} + \rho_{attr}}{2}, 0). \tag{9}$$

The adjacency matrix M is redefined using ρ_{uc} as follows:

$$M_{i,j} = \begin{cases} 1 & \text{if } (u_i, u_j) \in Y_S, \\ \rho_{uc} & \text{if } (u_i, u_j) \in Y_U, \\ 0 & else. \end{cases} \tag{10}$$

Using this updated adjacency matrix M, the CoFAD model performs fuzzy spectral clustering as described in Section 3.1. This integration enables CoFAD to incorporate feasibility-driven connections for unseen compositions, enhancing the clustering process and improving generalization to novel attribute-object pairs. For more details on the feasibility score, refer to the prior study [8].

4.3. label Adjustment

To represent the superordinate concepts of textual features, the centroid c^t_k for a cluster C_k is calculated as:

$$c^t_k = \frac{1}{|C_k|} \sum_{i \in C_k} x^t_i. \tag{11}$$

The superordinate concept feature c^t_k is then aggregated as a weighted sum of cluster centroids[1], using the fuzzy membership values computed earlier. The aggregated superordinate concept feature is defined as:

$$x^{super}_i = \sum_{k=1} u_{i,k} \cdot c^t_k, \tag{12}$$

where $u_{i,k}$ represents the fuzzy membership of label i in cluster k.

Finally, a label adjustment layer projects the concatenated concept features $[x^t_i, x^{super}_i]$ into an adjusted label feature x^{adj}_i. This adjustment enables the model to learn the relationships between primitives and multiple superordinate concepts effectively.

[1] To avoid confusion, the centroids μ_k derived in Section 4.1 are calculated from eigenvectors, whereas c^t_k is derived from text features.

 doi:10.20944/preprints202508.0451.v1

4.4. Training and Inference

Training Objectives. The logits for predicting the attribute and object labels of an image x are given by:

$$logit_a = cos(x_a^v, x_{a_i}^{adj}), \quad logit_o = cos(x_o^v, x_{o_j}^{adj}). \tag{13}$$

To encourage the model to consider the full search space $A \times O$, pairwise summations are computed as $logit_c = logit_a + logit_o$. The classification losses are defined as:

$$\begin{aligned} L_a &= CE(logit_a, y_a), \\ L_o &= CE(logit_o, y_o), \\ L_c &= CE(logit_c, y), \end{aligned} \tag{14}$$

where $CE(\cdot)$ denotes the Cross-Entropy objective function. The total loss L is a weighted sum of these losses:

$$L = \lambda_a L_a + \lambda_o L_o + \lambda_c L_c, \tag{15}$$

where $\lambda_a, \lambda_o, \lambda_c \in R$ are weights balancing the contributions of different losses.

Inference. During testing, the logits of attributes and objects are combined using a pairwise product to adjust the composition logits. The most likely composition is then predicted as:

$$\begin{aligned} \hat{p}(y_{i,j}|x) &= \omega(logit_c) + \omega(logit_a) \cdot \omega(logit_o), \\ \hat{y} &= argmax(\hat{p}(y_{i,j}|x)) \end{aligned} \tag{16}$$

where ω represents the softmax function.

5. Experiments

5.1. Experimental Setup

Dataset CoFAD was evaluated on three widely used CZSL benchmark datasets: MIT-States [16], UT-Zappos [17], and C-GQA [7].

- **MIT-States**, collected via older search engines, includes diverse compositions without distinguishing between living and non-living entities, such as "Burnt Wood" or "Tiny Cat." It contains 115 attributes and 245 objects, with 26,114 out of 28,175 compositions being non-existent labels (≈93%).
- **UT-Zappos** focuses on fine-grained images of shoes, such as "Suede Slippers" or "Cotton Sandals." It includes 16 attributes and 12 objects, with 76 out of 192 compositions being non-existent labels (≈40%).
- **C-GQA**, built on the Stanford GQA dataset [23], shares similar primitives with MIT-States but includes a significantly larger number of labels. It comprises 413 attributes and 674 objects, resulting in nearly 280,000 possible compositions. However, only 7,555 compositions are valid, with approximately 97% being non-existent pairs.

For a fair comparison, we used the datasets and train/validation/test splits provided by previous work [5,20].

5.2. Metrics

We adopt the established evaluation protocols [5,7,24] and report all results using four key metrics. Specifically, we measure the best seen score (**S**), where a large bias term limits the model to predicting only among seen labels, and the best unseen score (**U**), which reflects the model's zero-shot performance by predicting only unseen labels. To assess the balance between seen and unseen performance, we report the best harmonic mean (**HM**), which captures the trade-off between the two. Additionally, we provide the area under the seen-unseen curve (**AUC**) by varying the calibration bias. Both HM and AUC are core metrics for quantitatively evaluating models in CZSL tasks, offering comprehensive insights into their generalization capabilities.

 doi:10.20944/preprints202508.0451.v1

Table 2. OW-CZSL results on three benchmark datasets. The performance of baseline models is reported from their respective papers.

Models	MIT				UT-Zappos				C-GQA			
	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
CLIP [21]	30.1	14.3	12.8	3.0	15.7	20.6	11.2	2.2	7.5	4.6	4.0	0.3
CoOp [25]	34.6	9.3	12.3	2.8	52.1	31.5	28.9	13.2	21.0	4.6	5.5	0.7
PromptCompVL [26]	48.5	16.0	17.7	6.1	64.6	44.0	37.1	21.6	-	-	-	-
CSP [20]	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.2
HPL [27]	46.4	18.9	19.8	6.9	63.4	48.1	40.2	24.6	30.1	5.8	7.5	1.4
GIPCOL [28]	48.5	16.0	17.9	6.3	65.0	45.0	40.1	23.5	31.6	5.5	7.3	1.3
DFSP(i2t) [29]	47.2	18.2	19.1	6.7	64.3	53.8	41.2	26.4	35.6	6.5	9.0	2.0
DFSP(BiF) [29]	47.1	18.1	19.2	6.7	63.5	57.2	42.7	27.6	36.4	7.6	10.6	2.4
DFSP(t2i) [29]	47.5	18.5	19.3	6.8	66.8	60.0	44.0	30.3	38.3	7.2	10.4	2.4
PLID [30]	9.1	18.7	20.0	.3	**67.6**	55.5	46.6	30.8	39.1	7.5	10.6	2.5
Troika [14]	48.8	18.7	20.1	7.2	66.4	1.2	47.8	3.0	0.8	7.9	10.9	.7
CDS-CZSL [15]	**49.4**	**21.8**	**22.1**	**8.5**	64.7	**61.3**	8.2	32.3	37.6	.2	1.6	.7
CoFAD (Ours)	45.5	1.6	0.2	.3	7.4	59.7	**50.1**	**34.0**	**44.6**	**9.1**	**12.5**	**3.4**

5.3. Implementation Details

In our experiments, all baseline models, as well as our proposed model CoFAD, were implemented using PyTorch [31] and utilized the pre-trained CLIP ViT-L/14 as the backbone. The models were trained and evaluated on a single NVIDIA A5000 GPU and an Intel Xeon Silver 4314 Processor. For spectral clustering in CoFAD, the number of clusters was set to 4 for the UT-Zappos and MIT-States datasets, and 8 for the C-GQA dataset. The weights $\lambda_a, \lambda_o, \lambda_c$ were fixed at 1.0 across all settings. For further details, please refer to the appendix A.

5.4. Comparison with State-of-the-Arts

Table 2 presents a comparison of our method with previous studies in the Open-World setting. CoFAD demonstrates either superior performance or comparable results to previous state-of-the-art (SOTA) methods across all datasets. On both UT-Zappos and C-GQA, CoFAD demonstrates its strength in generalization, achieving highest performance in HM (49.9 on UT-Zappos and 12.5 on C-GQA) and AUC (33.9 on UT-Zappos and 3.4 on C-GQA). On MIT-States, CoFAD achieves an HM of 20.2 and an AUC of 7.3, which is slightly lower than CDS-CZSL [15] but achieved with a significantly reduced search space ($A + O$) compared to $A \times O$. This competitive performance and substantial improvement in cost efficiency underscores the effectiveness and strengths of CoFAD in addressing CZSL tasks. A detailed comparison of cost efficiency is provided in the following section.

5.5. Cost Efficiency

We conducted a comparative analysis of training cost efficiency against state-of-the-art (SOTA) models, as shown in Figure 3. The comparison was conducted using a compositionally diverse CGQA dataset in an open-world setting, with all models trained using a batch size of 2 and ViT-B/32 as the backbone architecture. The results demonstrate a dramatic improvement in both training time and GPU memory usage for our approach. Most high-performance CLIP-based models generate compositional features covering the $|A| \times |O|$ search space, leading to substantial GPU memory consumption and extended training duration. In contrast, the proposed CoFAD adopts a multi-label approach, reducing the search space to $|A| + |O|$. This results in GPU memory usage being reduced by at least 2× and up to 4× compared to baselines. Moreover, training time is reduced by at least 3× and up to 50×. Notably, despite CSP [20] being the smallest model among the baselines, CoFAD achieves over 2× improvements in both speed and memory efficiency while simultaneously delivering nearly double the performance.

 doi:10.20944/preprints202508.0451.v1

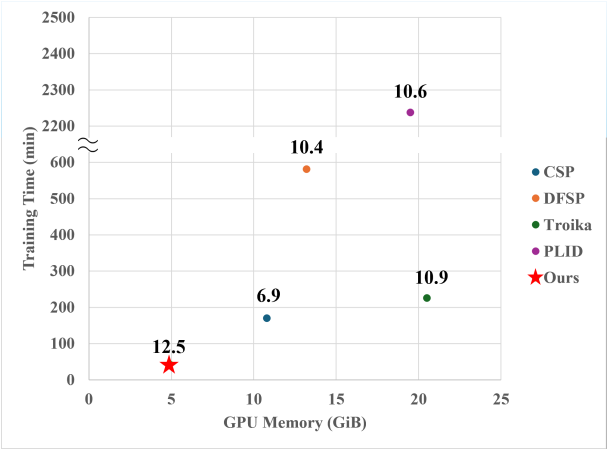

Figure 3. Comparison of the efficiency of our model and baseline methods on the C-GQA dataset. The numbers above each data point indicate the HM performance on C-GQA dataset as reported in Table 2.

5.6. Discussion

CoFAD utilizes a multi-label classification approach, enabling it to learn both primitives and the full compositions even in the closed-world setting. To address the challenge of handling a large search space, we conducted experiments using two methods: (1) a masking method, where logits for labels outside the closed-world pairs are multiplied by "$-1e8$" to mask them during loss computation, and (2) a discard method, where logits for these labels are excluded entirely, and their loss is not computed. Table 3 presents the experimental results in the closed-world setting. The masking method demonstrates improved performance and produces results comparable to those of SOTA models, whereas the discard method leads to a decrease in performance. This result suggests that CoFAD, which learns the relationships between primitives and their superordinate concepts, is negatively affected by the uncertainty introduced when unseen compositions are discarded in the loss function. By contrast, calculating the loss for all labels while using masking to identify irrelevant compositions allows the model to effectively learn and distinguish meaningful patterns, proving to be more effective for Zero-Shot Learning tasks.

Table 3. CW-CZSL results on two benchmark datasets. The performance of baseline models is reported from their respective papers.

Models	UT-Zappos				C-GQA			
	S	U	HM	AUC	S	U	HM	AUC
CSP	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
HPL	63.0	68.8	48.2	35.0	30.8	28.4	22.4	7.2
GIPCOL	65.0	68.5	48.8	36.2	31.9	28.4	22.5	7.1
DFSP(t2i)	66.7	71.7	47.2	36.0	38.2	32.0	27.1	10.5
PLID	67.3	68.8	52.4	38.7	38.8	33.0	27.9	11.0
Troika	66.8	73.8	**54.6**	**41.7**	41.0	**35.7**	**29.4**	**12.4**
CDS-CZSL	63.9	**74.8**	52.7	39.5	38.3	34.2	28.1	11.1
CoFAD	66.3	72.7	54.2	40.7	**45.4**	29.2	28.0	11.5
CoFAD$_{masking}$	**67.1**	71.6	54.2	41.1	44.6	29.5	28.6	11.5
CoFAD$_{discard}$	30.3	34.3	24.3	8.6	32.0	13.3	13.7	3.3

5.7. Qualitative Results

The visualization of contextual labels is challenging due to the complex interrelations among them, which makes analysis nontrivial [32]. Inspired by prior works [4,6,18,24,32], we perform a qualitative evaluation under scenarios where primitives are associated with different superordinate concepts, leading to significant visual differences.

 doi:10.20944/preprints202508.0451.v1

As illustrated in Figure 4, CoFAD consistently demonstrates superior contextual reasoning in distinguishing compositions. For the cases where two objects belong to different superordinate concepts but share the same attribute, CoFAD enables better semantic alignment by accurately identifying related compositions, such as "Caramelized Beef" or "Caramelized Sugar". Without CoFAD, the model fails to maintain contextual consistency, misclassifying labels as semantically unrelated combinations like "Caramelized Sauce" or "Molten Sugar". Similarly,for the cases where two attributes belong to different superordinate concepts but share the same object, CoFAD improves semantic coherence by correctly interpreting visual attributes, such as "Diced Cheese" and "Burnt Fence," whereas the absence of CoFAD leads to unrelated predictions like "Splintered Wood" or "Moldy Cheese."

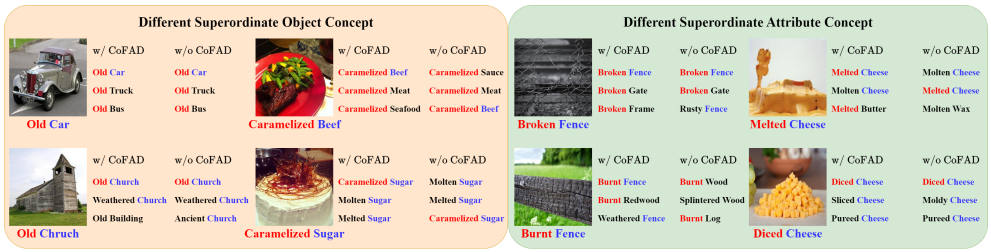

Figure 4. Qualitative results. We present the top-3 predictions for cases where objects share the same attribute but belong to different superordinate concepts, and where attributes share the same object but belong to different superordinate concepts, comparing the results with and without applying CoFAD. Red and blue indicate the ground truth attributes and objects, respectively.

Earlier, we noted that the multi-label approach tends to be biased in the training set when loosely related superordinate concepts share similar attributes, leading to poor recognition of features in unseen compositions. This bias is also evident in the qualitative results, where for food categories involving sugar and cheese, the model without CoFAD often leans toward the "Molten" attribute. In contrast, the results obtained by applying CoFAD demonstrate that this issue has been effectively mitigated.

6. Conclusions

This paper introduces a novel approach termed **C**oncept-**o**riented **F**eature **AD**justment (CoFAD), designed specifically for Compositional Zero-shot Learning (CZSL). CoFAD addresses the critical trade-off between performance and computational efficiency inherent in multi-label and multi-class classification approaches. By incorporating a design that enables the model to effectively capture and learn the relationships between individual labels and their corresponding superordinate concepts, CoFAD demonstrates remarkable generalization capabilities to unseen attribute-object compositions. Comprehensive experiments conducted on multiple benchmark datasets highlight the superior performance and efficiency of CoFAD. These findings emphasize the pivotal role of concept-oriented feature learning in real-world CZSL scenarios and establish a foundation for advancing efficient multi-label classification strategies in future research.

Author Contributions: Conceptualization, S.K. and Y.S.C.; methodology, S.K.; software, S.K.; validation, S.K.; formal analysis, S.K.; investigation, S.K.; resources, S.K.; data curation, S.K.; writing—original draft preparation, S.K.; writing—review and editing, S.K. and Y.S.C.; visualization, S.K.; supervision, Y.S.C.; project administration, Y.S.C.; funding acquisition, Y.S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant (No. 2018R1A5A7059549) and the Institute of Information and communications Technology Planning and evaluation (IITP) grant (No. RS-2020-II201373), funded by the Korean Government (MSIT: Ministry of Science and Information and Communication Technology).

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not applicable.

 doi:10.20944/preprints202508.0451.v1

Data Availability Statement: The MIT-States dataset can be downloaded from https://web.mit.edu/phillipi/Public/states_and_transformations/ (accessed on 3 March 2024). The UT-Zappos dataset can be downloaded from https://vision.cs.utexas.edu/projects/finegrained/utzap50k/ (accessed on 3 March 2024). The C-GQA dataset can be downloaded from https://s3.mlcloud.uni-tuebingen.de/czsl/cgqa-updated.zip (accessed on 3 March 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rosch, E.; Mervis, C.B.; Gray, W.D.; Johnson, D.M.; Boyes-Braem, P. Basic objects in natural categories. *Cognitive psychology* **1976**, *8*, 382–439.
2. Rosch, E. Principles of categorization. *Cognition and categorization/Erlbaum* **1978**.
3. Misra, I.; Gupta, A.; Hebert, M. From red wine to red tomato: Composition with context. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1792–1801.
4. Nagarajan, T.; Grauman, K. Attributes as operators: factorizing unseen attribute-object compositions. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 169–185.
5. Purushwalkam, S.; Nickel, M.; Gupta, A.; Ranzato, M. Task-driven modular networks for zero-shot compositional learning. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3593–3602.
6. Li, Y.L.; Xu, Y.; Mao, X.; Lu, C. Symmetry and group in attribute-object compositions. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11316–11325.
7. Naeem, M.F.; Xian, Y.; Tombari, F.; Akata, Z. Learning graph embeddings for compositional zero-shot learning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 953–962.
8. Mancini, M.; Naeem, M.F.; Xian, Y.; Akata, Z. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on pattern analysis and machine intelligence* **2022**, *46*, 1545–1560.
9. Karthik, S.; Mancini, M.; Akata, Z. Revisiting visual product for compositional zero-shot learning. In Proceedings of the NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications, 2021.
10. Kim, S.; Lee, S.; Choi, Y.S. Focusing on valid search space in Open-World Compositional Zero-Shot Learning by leveraging misleading answers. *IEEE Access* **2024**.
11. Anwaar, M.U.; Pan, Z.; Kleinsteuber, M. On leveraging variational graph embeddings for open world compositional zero-shot learning. In Proceedings of the Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 4645–4654.
12. Karthik, S.; Mancini, M.; Akata, Z. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9336–9345.
13. Li, X.; Yang, X.; Wei, K.; Deng, C.; Yang, M. Siamese contrastive embedding network for compositional zero-shot learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 9326–9335.
14. Huang, S.; Gong, B.; Feng, Y.; Zhang, M.; Lv, Y.; Wang, D. Troika: Multi-path cross-modal traction for compositional zero-shot learning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 24005–24014.
15. Li, Y.; Liu, Z.; Chen, H.; Yao, L. Context-based and Diversity-driven Specificity in Compositional Zero-Shot Learning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 17037–17046.
16. Isola, P.; Lim, J.J.; Adelson, E.H. Discovering states and transformations in image collections. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1383–1391.
17. Yu, A.; Grauman, K. Fine-grained visual comparisons with local learning. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 192–199.
18. Hao, S.; Han, K.; Wong, K.Y.K. Learning attention as disentangler for compositional zero-shot learning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15315–15324.
19. Mancini, M.; Naeem, M.F.; Xian, Y.; Akata, Z. Open world compositional zero-shot learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5222–5230.

 doi:10.20944/preprints202508.0451.v1

20. Nayak, N.V.; Yu, P.; Bach, S.H. Learning to compose soft prompts for compositional zero-shot learning. *arXiv preprint arXiv:2204.03574* **2022**.
21. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 8748–8763.
22. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In Proceedings of the International conference on machine learning. PMLR, 2019, pp. 2790–2799.
23. Hudson, D.A.; Manning, C.D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6700–6709.
24. Zhang, T.; Liang, K.; Du, R.; Sun, X.; Ma, Z.; Guo, J. Learning invariant visual representations for compositional zero-shot learning. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 339–355.
25. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision* **2022**, *130*, 2337–2348.
26. Xu, G.; Kordjamshidi, P.; Chai, J. Prompting large pre-trained vision-language models for compositional concept learning. *arXiv preprint arXiv:2211.05077* **2022**.
27. Wang, H.; Yang, M.; Wei, K.; Deng, C. Hierarchical prompt learning for compositional zero-shot recognition. In Proceedings of the Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 2023, pp. 1470–1478.
28. Xu, G.; Chai, J.; Kordjamshidi, P. GIPCOL: Graph-Injected Soft Prompting for Compositional Zero-Shot Learning. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 5774–5783.
29. Lu, X.; Guo, S.; Liu, Z.; Guo, J. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23560–23569.
30. Bao, W.; Chen, L.; Huang, H.; Kong, Y. Prompting language-informed distribution for compositional zero-shot learning. In Proceedings of the European Conference on Computer Vision. Springer, 2025, pp. 107–123.
31. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **2019**, *32*.
32. Saini, N.; Pham, K.; Shrivastava, A. Disentangling visual embeddings for attributes and objects. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13658–13667.