DMSD: Dual-Modal Semantic Disentanglement for Compositional Zero-Shot Learning

Pan Yang[1], Jing Yang[1]*, Xiaoli Ruan[1], Yuling Chen[1], Yuankai Wu[2], Quan Zhou[3], Xu Wang[1]
[1]Guizhou Provincial Laboratory of Big Data, Guizhou University,
[2]Sichuan University, [3]Nanjing University of Posts and Telecommunications
gs.pyang24@gzu.edu.cn; jyang23@gzu.edu.cn

Abstract

The core challenge of Compositional Zero-Shot Learning (CZSL) lies in learning representations of sub-concepts (attributes and objects) from seen compositions and recognizing unseen novel compositions. Most existing CZSL methods primarily focus on prompt optimization on the textual side, while overlooking insufficient visual attribute–object sub-concepts disentanglement under a text-centric paradigm. To this end, we propose *DMSD*, a ***D***ual-***M***odal ***S***emantic ***D***isentanglement framework that jointly models visual and textual information to achieve effective sub-concept disentanglement. Specifically, *DMSD* introduces a *Contextual Prompt Space*, enabling both visual and textual modalities to be modeled under unified contextual semantic representations, thereby enhancing their alignment at the latent semantic level. Moreover, we design *Visual Sub-concept Prototypes* that explicitly extract and model visual sub-concept features, improving the independence and discriminability of visual sub-concept representations. Furthermore, to achieve fine-grained alignment between visual and textual sub-concepts, we propose a *Class-Centroid Bridging Module* that guides class centroids toward the textual semantic space, thereby ensuring cross-modal semantic consistency. Extensive experiments on three benchmark datasets (MIT-States, UT-Zappos, and C-GQA) demonstrate that *DMSD* achieves state-of-the-art performance in both closed-world and open-world settings. Our code is available at https://github.com/ybyangjing/DMSD.

1 Introduction

Humans have a powerful compositional cognitive ability: when encountering a novel combination such as "blue banana", we can decompose it into known sub-concepts ("blue" and "banana") and rapidly recognize the entirely new combination. Inspired by this human capability, Compositional Zero-Shot Learning (CZSL) (Misra et al., 2017; Atzmon et al., 2020; Li et al., 2020) has emerged as a research paradigm that learns sub-concepts from seen attribute–object combinations and subsequently recognizes novel combinations that never appear during training.

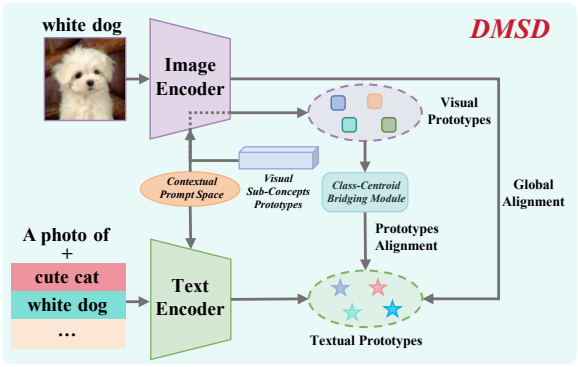

Figure 1: The overall pipeline of *DMSD*.

In recent years, large-scale Vision–Language Models (VLMs) have demonstrated remarkable capability in multimodal representation learning. VLMs such as CLIP (Radford et al., 2021) are trained on massive image–text pairs, where a text encoder projects textual inputs into a text space that is aligned with image representations obtained from an image encoder. Recent CZSL approaches built upon VLMs primarily focus on prompt tuning on the text side. Although these methods have achieved promising progress, they do not fundamentally resolve **the inherent limitation of performing attribute–object disentanglement in a text-centric manner.**

Specifically, prompt tuning methods improve the downstream generalization ability of large multimodal models by optimizing textual prompts for attributes and objects, as exemplified by approaches such as CoOp (Zhou et al., 2022b). However, such

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 30815–30826

text-centered disentanglement constructs new class prototypes solely within the textual semantic space, rather than effectively leveraging visual contextual information or achieving robust disentanglement of visual sub-concepts.

Most existing CLIP-based CZSL approaches follow this paradigm. The CSP (Nayak et al., 2023) framework first inserts learnable attribute and object embeddings into prompt templates (e.g., "a photo of [attribute] [object]"), requiring only minor modifications to the CLIP architecture. Benefiting from CLIP's prior knowledge, this approach achieves competitive performance. Subsequently, Troika (Huang et al., 2024) points out that CSP performs alignment only along a single path and fails to sufficiently encourage the disentanglement of sub-concepts, and therefore proposes multi-path vision–language alignment. Several later works (Wang et al., 2025; Jing et al., 2024) extend this paradigm. However, these methods attempt to directly disentangle attributes and objects within a global visual–semantic representation space. In essence, they force semantic components to be separated from holistic features, **lacking explicit disentanglement of visual sub-concepts.**

There is also a line of research (Bao et al., 2025a) that leverages large language models to enhance semantic representations. Such as, LOGICZSL (Wu et al., 2025) formalizes the relational knowledge embedded in large language models into logical rules and introduces a logic-guided loss, thereby explicitly modeling the semantic relationships among attributes, objects, and their compositions. However, such methods primarily rely on textual semantic knowledge to decouple sub-concept representations, **lacking a unified contextual prompt space** and making it difficult to effectively collaborate with visual information. They fail to sufficiently learn the class-centroids of sub-concepts.

Furthermore, since the sub-concept disentanglement process is mainly driven by prompts on the textual side, **the model is easily influenced by the prior semantic distribution and becomes biased.** For instance, if the co-occurrence probability of "red apple" is relatively high in the training corpus, the model may still predict "red apple" even when the apple in a test sample appears blue, continuing to follow the prior semantic bias and thus weakening its attention to real visual cues.

To address these challenges, we propose a ***D***ual-***M***odal ***S***emantic ***D***isentanglement framework (*DMSD*). By introducing a shared contextual prompt space together with visual sub-concept prototype prompts, *DMSD* effectively overcomes the limitations of text-centric disentanglement of visual sub-concepts. The overall framework is illustrated in Figure 1. For both the visual and textual branches, we construct a shared contextual prompt space that allows the two modalities to access consistent contextual information. In addition, we design prototypes to capture the semantic characteristics of visual sub-concepts, thereby enhancing their independent representations. Furthermore, to achieve fine-grained alignment of sub-concepts between the visual and textual modalities and reduce cross-modal discrepancy, we develop a class-prototype bridging module for *DMSD*. Built upon the backbone architecture of CLIP's text encoder, this module interacts with visual features layer-by-layer, connecting the visual space to the textual space and enabling effective representation of class centers across both modalities. In summary, the main contributions of our work are as follows:

- We propose *DMSD*, a multimodal prompt-learning framework for Compositional Zero-Shot Learning. By constructing a *Contextual Prompt Space*, *DMSD* enhances the collaborative modeling between the visual and textual modalities.
- Within this *Contextual Prompt Space*, we further design a set of *Visual Sub-concept Prototypes* to capture the fine-grained semantic characteristics of visual sub-concepts, thereby improving their independent and generalizable representations.
- We introduce a *Class-Centroid Bridging Module* for *DMSD*, which effectively reduces the representation gap between the visual and textual modalities and promotes cross-modal semantic alignment.
- Extensive experiments conducted on three benchmark CZSL datasets demonstrate that *DMSD* achieves SOTA in both closed-world and open-world settings, validating the effectiveness and superiority of the proposed approach.

2 Related work

2.1 Compositional Zero-Shot Learning

The goal of CZSL is to recognize unseen compositions of attributes and objects by leveraging

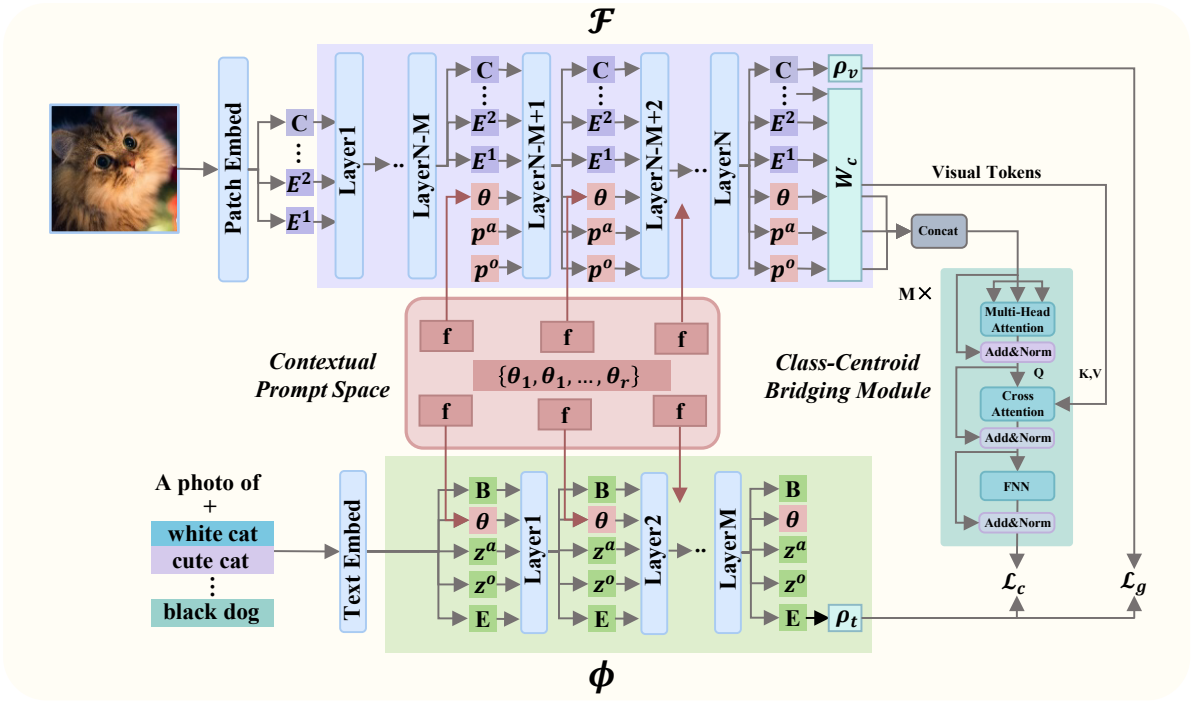

Figure 2: Overall framework of *DMSD*. "C" denotes the class token, "B" denotes the beginning-of-text token, and "E" denotes the end-of-text token. The set $\{\theta_1, \theta_2, \ldots, \theta_r\}$ represents the collection of prefix token embeddings. During training, contextual representations are introduced simultaneously into both the visual encoder and the text encoder, at the $(N-M)$-th layer and the first layer, respectively. In addition, a visual attribute prototype p^a and an object prototype p^o are incorporated.

prior knowledge about their seen attribute–object combinations. Early research on CZSL can be broadly grouped into two paradigms. The first paradigm of work jointly maps compositional text representations and image representations into a unified embedding space (Misra et al., 2017; Ma et al., 2024; Naeem et al., 2021; Ali Khan et al., 2023; Mancini et al., 2024; Nagarajan and Grauman, 2018). The second paradigm of work explicitly models attributes and objects separately by learning two independent classifiers (Yang et al., 2025; Li et al., 2022; Zhang et al., 2024, 2022; Saini et al., 2022; Li et al., 2023; Yang et al., 2020).

Currently, driven by the powerful multimodal representation capabilities of vision–language foundation models, a number of studies have adopted them as the backbone architecture for CZSL, forming a new research paradigm (Bao et al., 2025b; Xu et al., 2024; Nayak et al., 2023; Lu et al., 2023; Xu et al., 2022). For example, Troika (Huang et al., 2024) builds upon CLIP to jointly model attributes, objects, and their compositions, thereby learning their visual–language alignment relationships, and further refines the visual representations using Parameter-Efficient Transfer Learning (PETL) techniques (Houlsby et al., 2019). CDS-CZSL (Li et al., 2024) introduces a context-aware and diversity-guided specificity learning mechanism to place greater emphasis on attributes with richer semantic information. PLID (Bao et al., 2025b) leverages large language models to construct diverse and semantically expressive class distributions, and dynamically fuses predictions from the compositional space and the primitive space through a visual–language primitive decomposition module.

Despite the progress brought by these approaches, most existing methods still mainly focus on prompt engineering on the text side and the design of parameter-efficient adapters, while paying insufficient attention to multimodal collaborative modeling.

2.2 Prompt Learning in Vision Language Models

Prompt learning was first applied in natural language processing (Li and Liang, 2021) and has recently been extended to vision–language models. CoOp optimizes continuous text prompts for efficient few-shot transfer in CLIP, while Co-CoOp

(Zhou et al., 2022a) conditions prompts on images to improve generalization to novel categories (Zhou et al., 2022b).Recently, MaPLe (Khattak et al., 2023) argues that restricting prompt optimization to a single modality may limit adaptation capacity, and introduces learnable prompts in both visual and textual branches to enhance cross-modal alignment. MMRL (Guo and Gu, 2025a) and MMRL++ (Guo and Gu, 2025b) add a shared representation space at higher encoder layers for joint optimization of instance and class features. In CZSL task, CSP (Nayak et al., 2023) is the first to incorporate prompt learning into the vision–language model CLIP by treating attribute–object word embeddings as learnable parameters and optimizing their compositional prompt representations. Building on this, DFSP (Lu et al., 2023) introduces a cross-modal decomposition–fusion module that injects decomposed linguistic features into the visual feature learning process.

However, in the CZSL task, most existing methods primarily rely on text-centric attribute–object decoupling mechanisms, where such one-sided modeling fails to sufficiently disentangle visual sub-concepts. In contrast, our method explicitly decouples sub-concepts in the visual modality and further constructs a unified shared contextual prompt space for both the visual and textual branches, thereby enhancing cross-modal compositional modeling capability.

3 Fundamental Method

3.1 Task Formulation

In CZSL, each image x is assigned a label $c = (a, o)$ formed by combining one attribute a from the attribute set A and one object o from the object set O. The goal of CZSL is to leverage the visual representations of attributes and objects learned from the seen composition set C_s during training, so that the model can recognize attribute–object compositions $c \in C_u$ that do not appear in the training phase, where $C_s \cap C_u = \oslash$. CZSL can be divided into different settings. If the test composition set C_t is disjoint from the seen composition set C_s, i.e., $C_t \cap C_s = \oslash$, this is referred to as conventional CZSL. $C_t \cap C_s \neq \oslash$, the setting is called Closed-World CZSL (CW-CZSL). Furthermore, in the Open-World CZSL (OW-CZSL) setting, the test composition space covers all possible attribute–object pairs, i.e., $C_t \equiv A \times O$. Since $|C_t| \gg |C_s|$, the model must generalize from a very limited number of seen compositions to a large number of unseen ones, making this task particularly challenging.

3.2 Representations Encoding

We adopt the pretrained CLIP model with a Vision Transformer (Dosovitskiy et al., 2021) architecture as the backbone for both the image encoder $\mathcal{F}$ and the text encoder ϕ in *DMSD*.

Image Encoder The image encoder $\mathcal{F}$ consists of N Transformer layers. Given an input image x, it is divided into P patches, and each patch is mapped into the feature space to obtain the patch embedding $E_0 \in \mathbb{R}^{P \times d_v}$. A class token c_0 are then added to form the final input sequence. This sequence is subsequently updated layer by layer as follows:

$$[c_i, E_i] = \mathcal{F}_i([c_{i-1}, E_{i-1}]), i = 1, 2, ..., N. \quad (1)$$

After all Transformer layers, the projection matrix ρ_v maps the final class token output c_N into the shared vision–language space:

$$f = \rho_v(c_N). \quad (2)$$

We regard f as the global semantic representation of the image, denoted as $f^g \in \mathbb{R}^d$.

Text Encoder The text encoder ϕ consists of M Transformer layers. Following existing prompt tuning approaches for CZSL, we feed the attribute label set $A = \{a_i\}_{i=1}^n$ and the object label set $O = \{o_j\}_{j=1}^m$ into the pretrained CLIP word-embedding module to obtain the attribute embedding set E_A and object embedding set E_O:

$$E_A = \{z_i^a\}_{i=1}^n, \quad E_O = \{z_j^o\}_{i=1}^m, \quad (3)$$

where z^a and z^o denote the attribute and object embeddings, and n and m are the numbers of attribute and object labels in the dataset. For each candidate attribute, object, and attribute–object composition, we construct the corresponding soft prompts: $T_i^a = [\theta_1, \theta_2, ..., \theta_r, z_i^a], T_j^o = [\theta_1, \theta_2, ..., \theta_r, z_j^o]$, and $T_{i,j}^c = [\theta_1, \theta_2, ..., \theta_r, z_i^a, z_j^o]$, where $\{\theta_i\}_{i=1}^r$ are learnable prefix-token embeddings initialized with "a photo of". Let b_0 and c_0 denote the embeddings of the beginning-of-text and end-of-text tokens. The text-token embeddings, together with their positional encodings, are fed into the M Transformer layers of the text encoder, with the

encoding process defined as:

$$[b_i, T_i^*, e_i] = \phi_i([b_{i-1}, T_{i-1}^*, e_{i-1}]) \quad (4)$$
$$, i = 1, 2, ..., M,$$

where $*$ indicates the corresponding set (attribute, object, or composition). After the final Transformer layer, the end-of-text token output e_M is projected into the shared vision–language space:

$$t^* = \rho_t(e_M^*), \quad (5)$$

where ρ_t is the text-projection matrix and $*$ indicates the corresponding set.

4 *DMSD*: Dual-Modal Disentanglement

Through a systematic analysis of existing CZSL methods, we observe that most approaches are text-centric: **they focus on optimizing the embeddings of attributes and objects in the text modality to achieve decoupling between sub-concepts.** However, such a decoupling strategy that relies solely on the text modality has inherent limitations.

To address this issue, we propose a novel method termed Dual-Modal Semantic Disentanglement for Compositional Zero-Shot Learning (*DMSD*). As illustrated in Figure 2, the overall framework of *DMSD* consists of three main components: an image encoder $\mathcal{F}$, a text encoder ϕ and a *Class-Centroid Bridging Module*. Specifically, on the image side, the model introduces a *Contextual Prompt Space*, *Visual Sub-Concept Prototypes*, and visual features that jointly interact, thereby explicitly decoupling attribute and object semantics within the visual modality and strengthening class-centroid learning in the joint vision–language space. Moreover, the *Class-centroid Bridging Module* projects the learned visual sub-concept prompt representations into the text space, reducing the semantic gap between modalities and aligning the class centers towards the text space. On the text side, we adopt soft prompt learning to optimize text embeddings, enabling effective decoupling at the level of textual sub-concepts. Finally, the entire model is optimized using a cross-entropy loss.

4.1 *Contextual Prompt Space*

Most existing CZSL methods mainly focus on optimizing prompts within a single modality, without considering class-centroid learning across both the visual and textual modalities. To address this limitation, we introduce a *Contextual Prompt Space*, in which the learnable prefix-token embedding set $\theta = \{\theta_i\}_{i=1}^r$ serves as the contextual prompts that bridge the two modalities. In addition, a learnable linear mapping function f is adopted to facilitate cross-modal interaction. For the visual modality, the contextual prompts are defined as:

$$\theta^v = \{\theta_i^v\}_{i=1}^M, \quad \theta_i^v = f_i^v(\theta), \quad (6)$$

and for the text modality:

$$\theta^t = \{\theta_i^t\}_{i=1}^M, \quad \theta_i^t = f_i^t(\theta_i^v), \quad (7)$$

where θ_i^v denotes the contextual prompt token of the visual modality at the $(N - M + i)$-th Transformer layer, and θ_i^t denotes the contextual prompt token of the textual modality at the i-th Transformer layer.

In general, the dimensionality of visual features is higher than that of textual features. Therefore, we first project the contextual prompts into the visual semantic space for alignment and fusion, and then compress them back into the textual representation space. This "expand-then-compress" design uses the visual semantic space as an intermediary, thereby enhancing the ability of textual representations to capture visual features.

4.2 *Visual Sub-concept Prototypes*

To encourage the model to learn independent representations for visual Sub-concepts, we introduce two learnable *Visual Sub-Concept Prototypes*: the attribute visual concept prototype $p_0^a \in \mathbb{R}^{d_v}$ and the object visual concept prototype $p_0^o \in \mathbb{R}^{d_v}$ These tokens are injected into the image encoder at the $(N - M + 1)$-th layer and are processed together with the image patch embeddings in the subsequent layers. For the image encoder, the forward process is given by

$$[c_i, E_i] = \mathcal{F}_i([c_{i-1}, E_{i-1}]) \quad (8)$$
$$, i = 1, \ldots, N - M,$$
$$[c_i, E_i, S_1^v, p_i^a, p_i^o] = \mathcal{F}_i([c_{i-1}, E_{i-1}, \quad (9)$$
$$\theta_1^v, p_0^a, p_0^o]), i = N - M + 1,$$
$$[c_i, E_i, S_k^v, p_i^a, p_i^o] = \mathcal{F}_i([c_{i-1}, E_{i-1}, \quad (10)$$
$$\lambda S_{k-1}^v + (1 - \lambda)\theta_k^v, p_{k-1}^a, p_{k-1}^o]),$$
$$k = i - N + M, i = N - M + 2, \ldots, N,$$

where λ is the information-propagation coefficient, S_{k-1}^v denotes the output token from the previous

layer. For the text encoder, the process is formulated as:

$$[b_i, S_i^v, z_i^a, z_i^o, e_i] = \phi_i([b_{i-1}, \theta_i^t, z_{i-1}^a, z_{i-1}^o, e_{i-1}]), i = 1 \quad (11)$$

$$[b_i, S_i^t, z_i^a, z_i^o, e_i] = \phi_i([b_{i-1}, \lambda S_i^t + (1-\lambda)\theta_i^t, z_{i-1}^a, z_{i-1}^o, e_{i-1}]), i = 2, ..., M. \quad (12)$$

For clarity, only the case of composition soft prompts T^c is presented here; in practice, attribute and object soft prompts are processed simultaneously.

4.3 Class-Centroid Bridging Module

To enhance the fine-grained alignment between visual prompts (attributes and objects) and textual sub-concepts, and to promote the learning of a unified class center, we design a *Class-Centroid Bridging Module*. The details are as follows.

We adopt the CLIP text encoder as the backbone of the *Class-Centroid Bridging Module*. To preserve the pretrained knowledge, its parameters are frozen. Then, we introduce a learnable projection matrix W_c to project the encoded image embeddings

$$[c, E, S^v, p^a, p^o] = W_c([c_N, E_N, S_M^v, p_M^a, p_M^o]). \quad (13)$$

For each Transformer layer in *Class-Centroid Bridging Module*, we perform the following operations:

$$H = \text{Multi-HeadAttention}([S^v, p^a, p^o]), \quad (14)$$

$$H' = H + \text{Cross-Attention}(H, [c, E, S^v, p^a, p^o], [c, E, S^v, p^a, p^o]), \quad (15)$$

$$H'' = H' + \text{FNN}(\text{LayerNorm}(H')), \quad (16)$$

where Multi-HeadAttention denotes the multi-head attention mechanism (Vaswani et al., 2017), FNN is a feed-forward neural network composed of multilayer perceptrons, and Cross-Attention denotes the cross-attention mechanism (Vaswani et al., 2017). The parameters of the Multi-Head Attention and FNN are frozen, while the parameters of the CrossAttention are trainable. By retaining the pretrained knowledge of the CLIP text encoder, the model can explicitly model cross-modal interactions layer by layer, which helps concentrate the learning capacity on relational modeling and semantic alignment. We use the features obtained from the last layer as the compositional semantic representations f^c. Similarly, we can obtain attribute semantic representations f^a and object semantic representations f^o.

4.4 Training and Inference

By combining prompt representations and semantic representations from each branch, the label probabilities for the image's attribute branch a, object branch o, composition branch c, and global branch g can be computed separately as

$$p(z_* \mid x) = \frac{\exp(f^z \cdot t_*^z/\tau)}{\sum_{k=1}^{|Z|} \exp(f^z \cdot t_k^z/\tau)}, \quad z \in \{a, o, c, g\}, Z = \begin{cases} A, & z = a \\ O, & z = o \\ C_s, & z = c \\ C_s, & z = g \end{cases}, \quad (17)$$

where $*$ denotes the label index, and τ denotes a learnable temperature parameter.

Base Loss. The cross-entropy loss of each branch encourages the model to explicitly recognize the semantic role associated with that branch.

$$\mathcal{L}_\xi = -\frac{1}{|\chi|} \sum_{x \in \chi} \log\, p(\xi|x), \quad \xi \in \{a, o, c, g\}. \quad (18)$$

Decoupling Loss. Considering the strong entanglement between attribute and object concepts in images, we design a visual concept embedding decoupling loss. Specifically, we reduce the cosine similarity between the attribute semantics f^a and the object semantics f^o in order to achieve semantic decoupling in the embedding space. The loss is defined as:

$$\mathcal{L}_{de} = cos(f^a, f^o). \quad (19)$$

The overall training objective is formulated as:

$$\mathcal{L} = 0.2 \cdot \mathcal{L}_{de} + \sum \alpha_\xi \mathcal{L}_\xi, \quad \xi \in \{a, o, c, g\}, \quad (20)$$

where α_ξ denotes the loss weighting coefficient.

During inference, the input image x is fed into *DMSD* to obtain the attribute prediction score, object prediction score, global prediction score, and composition prediction score. These scores are then combined through a linear weighting scheme to produce the final prediction for each attribute–object pair:

$$s = \beta \cdot p(g_{i,j}|x) + (1-\beta) \cdot p(c_{i,j}|x) + p(a_i|x) \cdot p(o_j|x), \quad (21)$$

where β is a weighting coefficient that controls the relative contributions of the global and compositional branches. The compositions with the highest scores are taken as the model's final predictions.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate *DMSD* on three CZSL benchmark datasets, namely C-GQA (Mancini et al., 2024), UT-Zappos (Yu and Grauman, 2014), and MIT-States (Isola et al., 2015). Following the generalized evaluation protocol proposed by (Mancini et al., 2024), both seen and unseen attribute–object compositions are evaluated during testing. The training set contains only the seen compositions of all attributes and objects, while the validation and test sets consist of a mixture of seen and unseen compositions.

Metrics. Consistent with the standard evaluation protocol adopted in prior work (Nayak et al., 2023), we report the following metrics under both closed-world and open-world settings: the best accuracy on seen classes (S), the best accuracy on unseen classes (U), the best harmonic mean (HM), and the area under the curve (AUC). Among these metrics, AUC is regarded as the primary indicator, as it provides a more comprehensive evaluation of model performance. In addition, since the test set contains both seen and unseen compositions, the seen compositions act as competing labels and may significantly interfere with the recognition of unseen ones. To mitigate this inherent bias, a calibration bias term — ranging from $-\infty$ to $+\infty$ is introduced to balance the trade-off between seen and unseen accuracies. A positive bias favors predictions toward unseen compositions, whereas a negative bias encourages the model to lean toward seen compositions.

Implementation Details. We adopt the pre-trained CLIP ViT-L/14 model as the backbone for both the image and text encoders in *DMSD*. The image encoder is fine-tuned via AdapterFormer (Chen et al., 2022), with all components implemented in PyTorch. Training and evaluating are performed on a single NVIDIA RTX 6000 GPU. Additional details are provided in the supplementary material.

5.2 Main Results

We separately test closed-world and open-world performance of *DMSD* and compare it with SOTA methods using the same backbone (ViT-L/14) (Radford et al., 2021).

Closed-world results. Table 1 reports the results of *DMSD* and other state-of-the-art CZSL methods under the closed-world setting on three benchmark datasets: MIT-States, UT-Zappos, and C-GQA. We observe that *DMSD* achieves the best performance across all metrics on all three datasets. In terms of AUC, which serves as the primary metric for overall model evaluation, *DMSD* yields relative improvements of 3.4%, 0.9%, and 8.5% over the second-best method on MIT-States, UT-Zappos, and C-GQA, respectively. These results indicate that the proposed dual-modal decoupling strategy of *DMSD* exhibits advantages in the CZSL task.

Open-world results. Table 3 reports the results of *DMSD* in the open-world setting. We observe that, when transitioning from the closed world to the more realistic open world, the performance of various state-of-the-art methods degrades to different extents, especially on unseen compositions. Nevertheless, *DMSD* still achieves the best or second-best performance across the three benchmark datasets (MIT-States, UT-Zappos, and C-GQA) in the open-world scenario. Specifically, *DMSD*obtains AUC scores of 8.8%, 35.8%, and 4.0% on the three datasets, respectively, ranking first or second among all methods. Although our AUC on UT-Zappos is lower than that of LOGICZSL, our method attains higher unseen accuracy, and it also surpasses LOGICZSL in terms of AUC on the more challenging C-GQA dataset. These results indicate that *DMSD* exhibits stronger compositional generalization ability in the open-world setting.

5.3 Ablation Study

In the following section, we report experimental results of the ablation study for our framework under the closed-world setting. Additional ablation studies are provided in the supplementary material.

Ablation on Contextual Prompt Space. To verify whether the Contextual Prompt Space truly facilitates latent alignment across different modalities, we introduce contextual prompts only at the beginning of the visual and textual streams specifically, at layers $N-M+1$ on the visual side and the first layer on the textual side-while no additional shared contextual representations are injected into

Method	MIT-States				UT-Zappos				C-GQA			
	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
CSP (Nayak et al., 2023)	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
DFSP (Lu et al., 2023)	46.9	52.0	37.3	20.6	66.7	71.7	47.2	36.0	38.2	32.0	27.1	10.5
CAILA (Zheng et al., 2024)	51.0	53.9	39.9	23.4	67.8	74.0	57.0	44.1	43.9	38.5	32.7	14.8
CDS-CZSL (Li et al., 2024)	50.3	52.9	39.2	22.4	63.9	74.8	52.7	39.5	38.3	34.2	28.1	11.1
Troika (Huang et al., 2024)	49.0	53.0	39.3	22.1	66.8	73.8	54.6	41.7	41.0	35.7	29.4	12.4
PLID (Bao et al., 2025b)	49.7	52.4	39.0	22.1	67.3	68.8	52.4	38.7	38.8	33.0	27.9	11.0
RAPR (Jing et al., 2024)	50.0	53.3	39.2	22.5	69.4	72.8	56.5	44.5	45.6	36.0	32.0	14.4
MSCI (Wang et al., 2025)	50.2	53.4	39.9	22.8	67.4	75.5	**59.2**	45.8	42.4	38.2	31.7	14.2
LOGICZSL (Wu et al., 2025)	50.8	53.9	40.5	23.4	69.6	74.9	57.8	45.8	44.4	39.4	33.3	15.3
Ours	**52.8**	**54.4**	**41.5**	**24.2**	**69.9**	**77.3**	57.7	**46.2**	**45.8**	**41.5**	**34.2**	**16.6**

Table 1: Comparison results with SOTA methods in **Closed-World** setting on MIT-States, UT-Zappos, and C-GQA. "S", "U", "HM", and "AUC" stand for best Seen accuracy, best Unseen accuracy, best Harmonic Mean, and Area Under the Curve, respectively. The best results are in **bold**, and the second-best results are marked with an underline.

Method	MIT-States				UT-Zappos				C-GQA			
	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
CSP (Nayak et al., 2023)	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.2
DFSP (Lu et al., 2023)	47.5	18.5	19.3	6.8	66.8	60.0	44.0	30.3	38.3	7.2	10.4	2.4
CAILA (Zheng et al., 2024)	51.0	20.2	21.6	8.2	67.8	59.7	49.4	32.8	43.9	8.0	11.5	3.1
CDS-CZSL (Li et al., 2024)	49.4	**21.8**	22.1	8.5	64.7	61.3	48.2	32.3	37.6	8.2	11.6	2.7
Troika (Huang et al., 2024)	48.8	18.7	20.1	7.2	66.4	61.2	47.8	33.0	40.8	7.9	10.9	2.7
PLID (Bao et al., 2025b)	49.1	18.7	20.0	7.3	67.6	55.5	46.6	30.8	39.1	7.5	10.6	2.5
RAPR (Jing et al., 2024)	49.9	20.1	21.8	8.2	69.4	59.4	47.9	33.3	**45.5**	**11.2**	14.6	**4.4**
MSCI (Wang et al., 2025)	49.2	20.6	21.2	7.9	67.4	63.0	**53.2**	**37.3**	42.0	10.6	13.7	3.8
LOGICZSL (Wu et al., 2025)	50.7	21.4	**22.4**	8.7	**69.6**	63.7	50.8	36.2	43.7	9.3	12.6	3.4
***DMSD* (Ours)**	**51.5**	**21.8**	22.1	**8.8**	67.8	**66.9**	50.2	35.8	44.6	**11.2**	**15.1**	4.0

Table 2: Comparison results with SOTA methods in **Open-World** setting on MIT-States, UT-Zappos, and C-GQA.

Prompt Space	**MIT-States**				**UT-Zappos**			
	S	U	H	AUC	S	U	H	AUC
×	49.7	54.0	39.8	22.8	68.4	74.2	57.3	44.5
✓	**52.8**	**54.4**	**41.5**	**24.2**	**69.9**	**77.3**	**57.7**	**46.2**

Table 3: Ablation of Contextual Prompt Space.

Visual Prototypes	**MIT-States**				**UT-Zappos**			
	S	U	H	AUC	S	U	H	AUC
None	49.2	54.1	39.8	22.7	**70.5**	74.6	57.6	44.2
p^a	50.4	54.4	40.0	23.3	70.2	75.6	57.4	45.8
p^o	50.2	54.2	39.8	23.2	69.8	74.8	**57.9**	45.5
$p^a + p^o$	**52.8**	**54.4**	**41.5**	**24.2**	69.9	**77.3**	57.7	**46.2**

Table 4: Ablation on Visual Sub-Concept Prototypes.

subsequent layers of either modality. As shown in Table 3, when the shared space is enabled, the model performs significantly better than the variant without the shared space on both MIT-States and UT-Zappos, achieving relative improvements of 6.1% and 3.9% in AUC, respectively. This advantage mainly arises from the fact that the shared contextual space effectively preserves the latent alignment between modalities.

Ablation on Visual Sub-concept Prototypes. To verify the rationality of the proposed visual prototypes learning design, we conduct an ablation study by removing p^a and p^o to observe the corresponding performance changes. As shown in Table 4, when both p^a and p^o are used $(p^a + p^o)$, the model achieves the best performance on both datasets. Compared with the setting where all visual prompts are removed (None), the AUC im-

proves by 6.6% and 4.5% on MIT-States and UT-Zappos, respectively. These results indicate that our visual prompt design helps extract more discriminative visual sub-concept features and mitigates the negative effects introduced by purely text-centered disentanglement.

6 Conclusion

In this work, we investigate the inherent issues of text-centric attribute-object sub-concept disentanglement methods, which stem from the reliance solely on prompt tuning on the text side for sub-concept disentanglement and propose Dual-Modal Semantic Disentanglement (*DMSD*), which jointly aligns and disentangles visual and textual sub-concepts via a shared cross-modal prompt space, visual sub-concept prototypes, and a class center bridging module. Experiments show that *DMSD* consistently outperforms prior methods on multiple benchmarks in both closed- and open-world CZSL settings.

Limitations

Although *DMSD* alleviates text-centric disentanglement insufficiency by introducing *Contextual Prompt Space* and *Visual Sub-Concept Prototypes*, it still suffers from two limitations: (1) *Contextual Prompt Space* in the framework relies on a large number of representation alignment parameters, which may increase the risk of model overfitting; (2) Although *Class-Centroid Bridging Module* helps narrow gap between visual and textual modalities, it has not yet fully leveraged the representational potential of the visual modality itself.

Acknowledgments

This work is supported by the Key Project of Guizhou Basic Research Program (QKHZD [2026] 047), National Natural Science Foundation of China under Grant (Nos. 62441608, 62166005), Scientific and Technological Innovation Platform Research Project of Guizhou Province (CXP-TXM[2025]024), Science and Technology Research Project of Guizhou Provincial Department of Education (No.QJJ[2025]014), Guizhou University Science and Technology Innovation Team ([2024] No.07), Guizhou University Basic Research Fund ([2024]08).

References

Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, A. Pagani, Didier Stricker, and Muhammad Zeshan Afzal. 2023. Learning attention propagation for compositional zero-shot learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3817–3826.

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

Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. 2025a. Prompting language-informed distribution for compositional zero-shot learning. In *Computer Vision – ECCV 2024*, pages 107–123, Cham. Springer Nature Switzerland.

Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. 2025b. Prompting language-informed distribution for compositional zero-shot learning. In *Computer Vision – ECCV 2024*, pages 107–123, Cham. Springer Nature Switzerland.

Shoufa Chen, Chongjian GE, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, volume 35, pages 16664–16678. Curran Associates, Inc.

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. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuncheng Guo and Xiaodong Gu. 2025a. Mmrl: Multi-modal representation learning for vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25015–25025.

Yuncheng Guo and Xiaodong Gu. 2025b. Mmrl++: Parameter-efficient and interaction-aware representation learning for vision-language models. *Preprint*, arXiv:2505.10088.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Siteng Huang, Biao Gong, Yutong Feng, Min Zhang, Yiliang Lv, and Donglin Wang. 2024. Troika: Multi-path cross-modal traction for compositional zero-shot

learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24005–24014.

Phillip Isola, Joseph J. Lim, and Edward H. Adelson. 2015. Discovering states and transformations in image collections. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1383–1391.

Chenchen Jing, Yukun Li, Hao Chen, and Chunhua Shen. 2024. Retrieval-augmented primitive representations for compositional zero-shot learning. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. 2020. Symmetry and group in attribute-object compositions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11313–11322.

Yong-Lu Li, Yue Xu, Xinyu Xu, Xiaohan Mao, and Cewu Lu. 2022. Learning single/multi-attribute of object with symmetry and group. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9043–9055.

Yun Li, Zhe Liu, Hang Chen, and Lina Yao. 2024. Context-based and diversity-driven specificity in compositional zero-shot learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17037–17046.

Yun Li, Zhe Liu, Saurav Jha, and Lina Yao. 2023. Distilled reverse attention network for open-world compositional zero-shot learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1782–1791.

Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. 2023. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23560–23569.

Xingjiang Ma, Jing Yang, Jiacheng Lin, Zhenzhe Zheng, Shaobo Li, Bingqi Hu, and Xianghong Tang. 2024. Lvar-czsl: Learning visual attributes representation for compositional zero-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):13311–13323.

Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. 2024. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1545–1560.

Ishan Misra, Abhinav Gupta, and Martial Hebert. 2017. From red wine to red tomato: Composition with context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1160–1169.

Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. 2021. Learning graph embeddings for compositional zero-shot learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 953–962.

Tushar Nagarajan and Kristen Grauman. 2018. Attributes as operators: Factorizing unseen attribute-object compositions. In *Computer Vision – ECCV 2018*, pages 172–190, Cham. Springer International Publishing.

Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. 2023. Learning to compose soft prompts for compositional zero-shot learning. *Preprint*, arXiv:2204.03574.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Nirat Saini, Khoi Pham, and Abhinav Shrivastava. 2022. Disentangling visual embeddings for attributes and objects. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13648–13657.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Yue Wang, Shuai Xu, Xuelin Zhu, and Yicong Li. 2025. Msci: Addressing clip's inherent limitations for compositional zero-shot learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 2009–2017. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Peng Wu, Xiankai Lu, Hao Hu, Yongqin Xian, Jianbing Shen, and Wenguan Wang. 2025. Logiczsl: Exploring logic-induced representation for compositional zero-shot learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30301–30311.

Guangyue Xu, Joyce Chai, and Parisa Kordjamshidi. 2024. Gipcol: Graph-injected soft prompting for compositional zero-shot learning. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5762–5771.

Guangyue Xu, Parisa Kordjamshidi, and Joyce Chai. 2022. Prompting large pre-trained vision-language models for compositional concept learning. *Preprint*, arXiv:2211.05077.

Jing Yang, Xingjiang Ma, Yuankai Wu, Chengjiang Li, Zhidong Su, Ji Xu, and Yixiong Feng. 2025. Aognczsl: An attribute- and object-guided network for compositional zero-shot learning. *Information Fusion*, 120:103096.

Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. 2020. Learning unseen concepts via hierarchical decomposition and composition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10245–10253.

Aron Yu and Kristen Grauman. 2014. Fine-grained visual comparisons with local learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199.

Tian Zhang, Kongming Liang, Ruoyi Du, Xian Sun, Zhanyu Ma, and Jun Guo. 2022. Learning invariant visual representations fornbsp;compositional zero-shot learning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, page 339–355, Berlin, Heidelberg. Springer-Verlag.

Yang Zhang, Songhe Feng, and Jiazheng Yuan. 2024. Continual compositional zero-shot learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 1724–1732. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zhaoheng Zheng, Haidong Zhu, and Ram Nevatia. 2024. Caila: Concept-aware intra-layer adapters for compositional zero-shot learning. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1710–1720.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16795–16804.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *Int. J. Comput. Vision*, 130(9):2337–2348.

A Implementation Details

In *DMSD*, we employ the pre-trained and frozen CLIP ViT-L/14 model as the backbone for both the image and text encoders, while the image encoder is fine-tuned using LoRA. Training and evaluation are performed on a single NVIDIA RTX 6000 GPU with 48 GB of memory. All components are implemented in PyTorch. Table 5 summarizes the hyperparameters used for the three datasets.

Hyper parameters	**MIT-States**	**UT-Zappos**	**C-GQA**
BatchSize	64	64	64
Learning Rate	10^{-4}	2.5×10^{-4}	10^{-4}
Epochs	10	10	10
Scheduler	StepLR	StepLR	StepLR
Weight Decay	10^{-5}	10^{-5}	10^{-5}
Optimizer	Adam	Adam	Adam
β	0.85	0.65	0.85
λ	0.8	0.8	0.8
α_a,α_o,α_c,α_g	0.5,0.5,1,1	0.5,0.5,1,1	0.5,0.5,1,1
Attribute Dropout	0.3	0.3	0.3

Table 5: Hyperparameters for MIT-States,UT-Zappos,C-GQA.

B Dataset Details

We evaluated *DMSD*on three CZSL benchmark datasets: MIT-States, UT-Zappos and C-GQA.

C-GQA Dataset (Mancini et al., 2024). With 39,298 images, this dataset offers extensive coverage of real-world concepts through a diverse range of objects and attributes. It is annotated for 7,767 distinct attribute-object pairs, establishing it as the largest benchmark in the field of Compositional Zero-Shot Learning (CZSL).

UT-Zappos Dataset (Yu and Grauman, 2014). This dataset contains 50,025 images of footwear, designed for fine-grained recognition across 12 shoe categories and 16 material attributes. A primary challenge it presents is the subtle visual distinction between materials like "synthetic leather" and "genuine leather," demanding models capable of discerning fine texture and material details.

MIT-States Dataset (Isola et al., 2015). Comprising 53,753 images of everyday scenes, this dataset includes 245 objects and 115 attributes, resulting in 1,962 unique compositions. The data was largely gathered automatically via early image search engines, with limited manual curation. Significant label noise in this dataset places a premium on model robustness.

The detailed dataset splits for each benchmark are reported in Table 6.

Dataset	A	O	Training		Validation			Test		
			C_s	I	C_s	C_u	I	C_s	C_u	I
C-GQA	413	674	5592	26k	1252	1040	7k	888	923	5k
UT-Zap50K	16	12	83	23k	15	15	3k	18	18	3k
MIT-States	115	245	1262	30k	300	300	10k	400	400	13k

Table 6: Statistics of datasets for Training / Validation / Test.

C Additional Ablation Study

Ablation on Class-Centroid Bridging Module. In Table 7, to investigate whether the class-centroid bridging module truly promotes the learning of unified class centers, we conducted ablation studies using three strategies: the first is the averaging strategy (mean), which only computes the mean of the features of $[O^v, p^a, p^o]$ to explore the alignment effect between the features extracted from the visual modality itself and the text features; the second is the text-encoder-only strategy (w/o CA), where we remove the Cross-Attention module from the class-centroid bridging module, relying solely on CLIP's pre-trained text encoder; the third is the strategy using the class-centroid bridging module (w/ CA). As shown in Table 4, when employing the class-centroid bridging module (w/ CA), the model achieves the best performance on both datasets. Compared to the averaging strategy (mean), the AUC on MIT-States and UT-Zappos shows relative improvements of 5.2% and 2.9%, respectively, indicating that our class-centroid bridging module helps reduce the gap between modalities and effectively projects visual semantics into the text space. Additionally, we found that the text-encoder-only strategy (w/o CA) underperforms compared to the averaging strategy. We hypothesize that relying solely on text encoding to bridge the text space leads to the loss of information inherent in the visual features, resulting in decreased performance.

Strategies	MIT-States				UT-Zappos			
	S	U	H	AUC	S	U	H	AUC
mean	51.4	52.9	40.0	23.0	68.3	76.9	56.2	44.9
w/o CA	50.3	53.1	39.3	22.7	68.9	76.1	56.8	45.1
w/ CA	**52.8**	**54.4**	**41.5**	**24.2**	**69.9**	**77.3**	**57.7**	**46.2**

Table 7: Ablation on class-centroid Bridging Module.

Ablation on Each Loss. In the closed-world setting, we conducted ablation experiments on the MIT-States and UT-Zappos datasets to validate the effectiveness of each loss function design. As shown in Table 8, the results indicate that the loss functions in *DMSD* can work synergistically, mutually enhancing each other and improving the overall model performance.

The results show that when using only $\mathcal{L}_g$, although CLIP can leverage prior knowledge to achieve relatively reasonable performance on MIT-States, it performs poorly on fine-grained tasks and shows limited effectiveness on the UT-Zappos dataset.

By introducing the $(\mathcal{L}_a + \mathcal{L}_o)$ loss functions, the model is able to extract visual features with more fine-grained semantics, leading to significant performance improvement on UT-Zappos. Furthermore, combining $\mathcal{L}_a + \mathcal{L}_o + \mathcal{L}_c + \mathcal{L}_{de}$ with $\mathcal{L}_g$ results in substantial performance gains on both MIT-States and UT-Zappos. Compared with using only $\mathcal{L}_g$, the proposed loss function design achieves relative AUC improvements of 12.6% and 8.5% on MIT-States and UT-Zappos, respectively.

Loss	MIT-States				UT-Zappos			
	S	U	HM	AUC	S	U	HM	AUC
$\mathcal{L}_g$	48.2	53.0	38.3	21.5	67.2	73.3	55.8	42.6
$\mathcal{L}_{a,o}$	41.6	52.0	34.8	18.0	66.5	72.1	54.0	40.6
$\mathcal{L}_{c,a,o}$	45.1	52.3	36.3	19.8	67.9	74.6	56.1	43.9
$\mathcal{L}_{g,a,o}$	48.5	53.3	38.6	21.9	68.3	76.9	56.2	44.9
$\mathcal{L}_{c,g,a,o}$	51.0	53.2	39.5	23.2	69.2	76.1	56.5	45.2
$\mathcal{L}_{c,g,a,o,de}$	**52.8**	**54.4**	**41.5**	**24.2**	**69.9**	**77.3**	**57.7**	**46.2**

Table 8: Ablation on Each Loss.