Prototype-Conditioned Imagination for Compositional Zero-Shot Learning

Yifan Zhu and Haofeng Zhang(✉)

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
yifanzhu12138@njust.edu.cn zhanghf@njust.edu.cn

Abstract. Compositional Zero-Shot Learning (CZSL) requires learning to compose primitive concepts from seen data and generalize to unseen attribute–object pairs. Existing methods typically disentangle visual features into isolated primitive components and align them with their corresponding text concepts or full compositions. However, they often suffer from ambiguous vision–language alignment and lack mechanisms to explicitly capture cross-primitive semantic interactions, ultimately hindering generalization to unseen compositions. To overcome these challenges, we introduce Prototype-Conditioned Imagination (PCI), a two-stage framework. In the first stage, Bidirectional Alignment and Disentanglement (BAD) couples cross-modal attention with entropic optimal-transport regularization to effectively align visual and textual representations while enforcing global disentanglement of attribute and object primitives, thereby mitigating ambiguous vision-language matching. In the second stage, Prototype-driven Compositional Modulation (PCM) explicitly models how attribute semantics modulate object features to synthesize compositional representations, making cross-primitive relations explicit and substantially improving recognition of unseen compositions. Evaluated under both closed-world and open-world protocols on four popular benchmarks, our approach achieves state-of-the-art performance, demonstrating superior generalization capabilities to unseen compositions. Our code is available at `https://github.com/YFan-Z1/PCI`.

Keywords: Compositional Zero-Shot Learning · Optimal Transport · Prototype Learning

1 Introduction

Compositional Zero-Shot Learning (CZSL) [36,18,6,23,11,12] studies how to recognize unseen attribute–object compositions by recombining primitive concepts learned from seen data. For example, humans can easily infer the concept of a "`sliced banana`" after learning "`yellow banana`" and "`sliced apple`." Early studies typically treated each attribute–object pair as an independent category [4] or attempted to disentangle visual factors through spatial decomposition and bilinear embeddings [36,18]. With the emergence of large-scale vision–language models (VLMs) [29,16,40] such as CLIP [29], recent research in CZSL has shifted

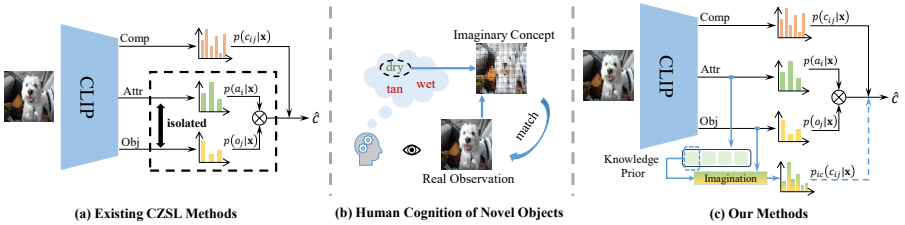

Fig. 1. Comparison of PCI with prior CZSL paradigms. Existing methods (a) often align attributes and objects with independent primitive branches, leaving cross-primitive interactions implicit and weakening generalization to unseen compositions. Inspired by human concept formation (b), PCI (c) synthesizes virtual compositions by using attribute knowledge to condition attribute prototypes on object context, making cross-primitive dependence explicit.

toward leveraging pretrained multimodal representations for robust cross-modal alignment. Huang et al. [7] introduced a multi-path framework that jointly models attributes, objects, and their compositions, while Wu et al. [33] employed a logic-induced learning paradigm to explore deeper semantic dependencies between primitives and compositions.

Although the aforementioned methods have achieved remarkable progress, they still suffer from two major limitations. 1) Ambiguous multimodal alignment. Despite the impressive cross-modal reasoning ability of VLMs, they often struggle with visually or semantically confusing image–text pairs, a problem that becomes more pronounced when attributes and objects must be recognized independently. For instance, in the C-GQA [24] dataset, attributes such as tan, brown, and beige all refer to similar brownish hues; in such cases, fine-grained visual textures are indispensable for enlarging the representational margin. In contrast, when distinguishing more abstract semantics, the textual prior itself acts as the dominant discriminative factor. 2) Lack of conditional interaction between attribute and object pathways. Existing CZSL models [5,7,32,33] commonly adopt multi-path architectures where attributes, objects, and compositions are trained independently. While this design allows each branch to make separate decisions, similar to the human process of first identifying individual attributes and objects before combining them, it overlooks the explicit conditional dependency between the primitives, thereby limiting the model's ability to capture the semantics inherent in compositional structures.

Learning fully disentangled visual–language representations and performing compositional modeling under conditional inference are crucial. However, disentangling attribute and object representations and then explicitly composing them remains challenging. We introduce Prototype-Conditioned Imagination (PCI), a two-stage framework that addresses these requirements sequentially. In the first stage, Bidirectional Alignment and Disentanglement (BAD) employs bidirectional cross-modal attention to produce deep, pathway-wise alignment, and incorporates an optimal-transport regularizer to suppress ambiguous

matches and enforce global consistency of the attention maps. In the second stage, Prototype-driven Compositional Modulation (PCM) performs compositional semantic exploration in the already-disentangled representation space via a conditional imagination mechanism that composes isolated attribute and object pathways. As illustrated in Fig. 1, PCM simulates the human process for hypothesizing novel object concepts: when faced with a novel object, humans naturally search a repertoire of known attributes, apply selected attributes to the object concept, and then judge whether the "imagined" composite concept matches the real observation.

The main contributions are summarized as follows.

- We propose Prototype-Conditioned Imagination for CZSL, a two-stage framework that achieves fine-grained disentanglement of the primitive representations and enables explicit, conditional attribute–object composition.
- We introduce **Bidirectional Alignment and Disentanglement**, which strengthens vision–language interaction via bidirectional attention and mitigates semantic ambiguity through an optimal-transport regularizer, yielding more globally consistent cross-modal alignment.
- We propose **Prototype-driven Compositional Modulation**, a mechanism that emulates human imaginative reasoning. PCM learns a set of dynamic prototypes and forms compositional associations from isolated primitive representations, effectively "imagining" attribute–object pairs and significantly enhancing generalization.
- Comprehensive experiments on four benchmarks establish new state-of-the-art results, corroborating the effectiveness of BAD and PCM.

2 Related Works

Compositional Zero-Shot Learning. CZSL aims to infer novel attribute–object combinations by learning shared semantics from existing primitives. Early approaches typically modeled attributes and objects independently, either by performing separate classification [1,17,23,27] or propagating information through graph convolutional networks [21,24] within a joint embedding space. With the rise of pretrained VLMs [29,40], recent studies [34,25,2,7,35] have shifted toward leveraging their multimodal representations to compose unseen concepts. Troika [7] introduces separate modeling branches for attributes, objects, and their compositions to fully exploit the pretrained knowledge of VLMs. MSCI [32] enhances fine-grained discrimination by utilizing intermediate CLIP features, while Logic [33] integrates relation knowledge distilled from large language models (LLMs) to uncover semantic connections between attributes and objects.

Learning compositional features. Learning compositional representations is central to CZSL, as it supports structured reasoning and transfer to unseen attribute–object pairs. Early lines of work [21,20,15,23] largely construct joint embedding spaces by concatenating primitive word vectors or rely on graph-based propagation to diffuse signals across composition nodes. However, such

designs insufficiently exploit primitive sharing across classes, often yielding suboptimal transfer to novel compositions. More recent approaches pursue conditional interactions among primitives to better capture how attributes and objects combine. For instance, CANet [31] learns attribute embeddings conditioned on both the object category and the input image, thereby adapting attribute semantics to visual context; IMAX [10] operates in the complex domain, introducing imaginary-connected compositional embeddings that encode pair-specific dependencies via phase.

In contrast to prior baselines, we introduce Prototype-driven Compositional Modulation, a primitive interaction mechanism inspired by human compositional cognition. When encountering an unfamiliar object, human cognition does not process attributes and categories in strict isolation, nor does it rely on the heuristic concatenation of observed cues. Instead, humans retrieve the most relevant concepts from memory based on perceived attributes, synthesizing plausible virtual representations through cognitive extrapolation to support robust recognition. PCM follows this intuition by learning dynamic prototypes that carry attribute semantics and modeling the "imagination" process as prototype-guided modulation applied to object features. This design makes cross-primitive interactions explicit and improves generalization to unseen compositions.

3 Method

3.1 Task Formulation

Let the attribute vocabulary be $\mathcal{A} = \{a_1, \ldots, a_M\}$ and the object vocabulary be $\mathcal{O} = \{o_1, \ldots, o_K\}$. We define the compositional label space as the Cartesian product $\mathcal{C} = \mathcal{A} \times \mathcal{O}$, which is partitioned into two mutually exclusive subsets: the seen compositions $\mathcal{C}_s$ and the unseen compositions $\mathcal{C}_u$, satisfying $\mathcal{C}_s \cap \mathcal{C}_u = \varnothing$. During training, the model only observes image–label pairs drawn from the seen subset, i.e., the training set is $\mathcal{T} = \{(\mathbf{x}, c) \mid \mathbf{x} \in \mathcal{X},\ c \in \mathcal{C}_s\}$, where $\mathcal{X}$ denotes the visual domain. Under the *Closed-world* evaluation protocol, the candidate labels at test time are restricted to the benchmark split, so the test time composition set is $\mathcal{C}_t = \mathcal{C}_s \cup \mathcal{C}_u$, whereas under the *Open-world* protocol, every possible attribute–object pairing is considered at test time, i.e., $\mathcal{C}_t = \mathcal{C}$.

3.2 Overview

PCI is a two-stage framework built upon CLIP to leverage rich multimodal priors. In CZSL, generalization to unseen compositions is hindered by ambiguous cross-modal primitive matching and weak primitive interaction; fully disentangling primitives while modeling their interaction is non-trivial. To this end, Stage I—Bidirectional Alignment and Disentanglement (BAD) performs strong vision–language alignment and intra-primitive disentanglement, yielding expressive, decoupled primitive representations. Stage II—Prototype-driven Compositional Modulation (PCM) builds on BAD, explicitly modeling cross-primitive interactions via prototype-driven modulation to improve recognition of unseen compositions. An overview of PCI is provided in Fig. 2.

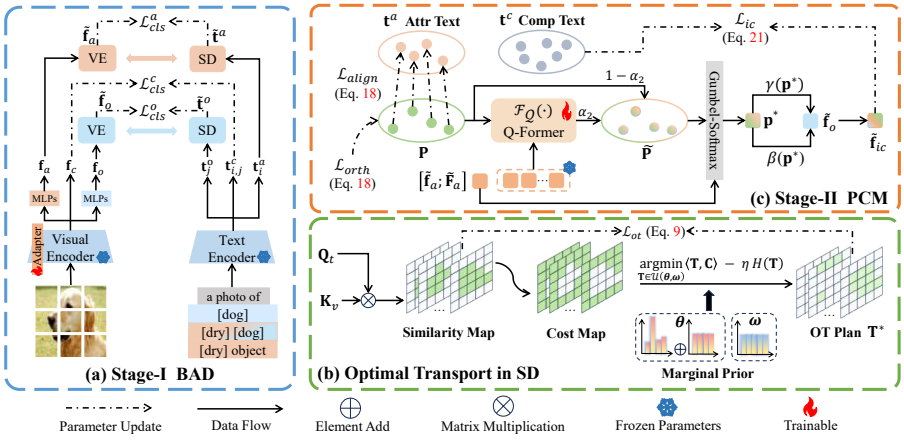

Fig. 2. (a) Stage-I BAD adopts a three-branch architecture to disentangle attribute, object, and composition semantics. Visual Enrichment (VE) enhances visual cues, while Semantic Disentanglement (SD) adaptively aligns visual evidence with textual tokens. (b) SD further leverages Optimal Transport to enforce text-guided attention competition over shared regions, yielding globally consistent alignment under marginal priors. (c) Stage-II PCM builds on BAD by learning an attribute prototype bank and modulating object features to generate compositional representations.

3.3 Bidirectional Alignment and Disentanglement

Feature Extraction. CLIP image encoder E_v serves as the visual backbone for an input image $\mathbf{x}$. The encoder outputs a set of patch tokens $\mathbf{F} \in \mathbb{R}^{N\times D}$ and a global token $\mathbf{f}_{\text{cls}} \in \mathbb{R}^D$, where N denotes the number of patch tokens; the latter is used as the composition-branch image feature, i.e., $\mathbf{f}_c := \mathbf{f}_{\text{cls}}$. Following the three-path paradigm [7], two independent MLP heads take $\mathbf{f}_{\text{cls}}$ as input and produce the attribute-oriented and object-oriented visual features, denoted by $\mathbf{f}_a$ and $\mathbf{f}_o$, respectively.

For textual descriptors, we adopt a soft, learnable prompting scheme [25,7] and a frozen CLIP text encoder E_t to produce pathway-specific text embeddings: for each attribute–object pair $c_{ij} = \langle a_i, o_j \rangle$, we instantiate an attribute prompt $\mathbf{R}_i^a = [\mathbf{r}_a^1, \ldots, \mathbf{r}_a^{l_{\text{ctx}}}, \mathbf{e}_i^a]$, an object prompt $\mathbf{R}_j^o = [\mathbf{r}_o^1, \ldots, \mathbf{r}_o^{l_{\text{ctx}}}, \mathbf{e}_j^o]$ and a composition prompt $\mathbf{R}_{i,j}^c = [\mathbf{r}_c^1, \ldots, \mathbf{r}_c^{l_{\text{ctx}}}, \mathbf{e}_i^a, \mathbf{e}_j^o]$, where the prefix context vectors $\{\mathbf{r}_*^k\}_{k=1}^{l_{\text{ctx}}}$ are learnable and $\mathbf{e}_i^a, \mathbf{e}_j^o$ are trainable vocabulary tokens for a_i and o_j; these prompted sequences are subsequently mapped through the text encoder to yield global, ℓ_2-normalized text representations:

$$\mathbf{t}_i^a = E_t(\mathbf{R}_i^a), \quad \mathbf{t}_j^o = E_t(\mathbf{R}_j^o), \quad \mathbf{t}_{ij}^c = E_t(\mathbf{R}_{i,j}^c). \tag{1}$$

Visual Enrichment. To achieve discriminative and semantically grounded representations, we enhance both the attribute and object pathways through a visual enrichment process. Let $\mathbf{f}_p \in \mathbb{R}^D$ denote the pathway-specific global feature with $p \in \{a, o\}$ (i.e., $\mathbf{f}_a$ for attributes and $\mathbf{f}_o$ for objects), and let the patch token

from the encoder be $\mathbf{F}$. A unified visual token sequence is formed by concatenation along the token dimension: $\mathbf{F}_p^{\text{uni}} = [\mathbf{f}_p; \mathbf{F}]$.

For each pathway, we treat the concatenated visual descriptor as the query, while the corresponding text embedding $\mathbf{t}^p$ (attribute text $\mathbf{t}^a \in \mathbb{R}^{|\mathcal{A}| \times D}$ or object text $\mathbf{t}^o \in \mathbb{R}^{|\mathcal{O}| \times D}$) serves as the key and value. The interaction is computed via scaled dot-product attention:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D}}\right)\mathbf{V}. \tag{2}$$

To preserve the original visual signal and stabilize training, the cross-attention is followed by residual connections and a feed-forward network (FFN):

$$\mathbf{z}_p^{\text{uni}} = \mathbf{F}_p^{\text{uni}} + \text{Attn}\big(\mathbf{Q} = \mathbf{F}_p^{\text{uni}},\, \mathbf{K} = \mathbf{t}^p,\, \mathbf{V} = \mathbf{t}^p\big), \tag{3}$$

$$\widetilde{\mathbf{F}}_p^{\text{uni}} = \mathbf{z}_p^{\text{uni}} + \text{FFN}\big(\text{LayerNorm}(\mathbf{z}_p^{\text{uni}})\big). \tag{4}$$

This cross-modal attention with residual pathways effectively injects linguistic priors into the visual tokens, allowing the model to emphasize regions and patterns most relevant to a given attribute or object. As a result, both branches benefit from text-guided semantic aggregation, forming richer and more discriminative representations that facilitate subsequent OT-guided disentanglement.

Semantic Disentanglement. Following VE, each pathway yields an enhanced visual descriptor $\widetilde{\mathbf{F}}_p^{\text{uni}} = [\widetilde{\mathbf{f}}_p; \widetilde{\mathbf{F}}_p]$ for $p \in \{a, o\}$. Prior work [7,32] typically applies vanilla attention to adjust static text embeddings for multimodal alignment. However, in the presence of semantically proximate vocabularies, unconstrained attention distributions tend to collapse onto identical high-saliency visual patches, engendering homogeneous and heavily entangled primitive correspondences. To address this, SD employs an entropic optimal transport (OT) framework [14,3] to enforce globally consistent, competitive evidence allocation. Before applying OT, a patch-normalized affinity is constructed by treating the pathway text embeddings as queries and the enhanced patch tokens as keys/values, i.e., $\mathbf{Q}_t = \mathbf{t}^p$ and $\mathbf{K}_v = \mathbf{V}_v = \widetilde{\mathbf{F}}_p$:

$$\mathbf{S} \;=\; \text{Softmax}\left(\frac{\mathbf{Q}_t \mathbf{K}_v^\top}{\sqrt{D}}\right). \tag{5}$$

Define the nonnegative cost matrix element-wise as $C_{ij} = 1 - S_{ij}$, where higher affinities in $\mathbf{S}$ incur lower transport costs. Given text- and patch-side marginals $\boldsymbol{\theta} \in \Delta^L$ (with $L \in \{|\mathcal{A}|, |\mathcal{O}|\}$) and $\boldsymbol{\omega} \in \Delta^N$, the entropic optimal transport plan is obtained by solving the following convex program:

$$\mathbf{T}^\star = \underset{\mathbf{T} \in \mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\omega})}{\arg\min} \; \langle \mathbf{T}, \mathbf{C} \rangle \;-\; \eta\, H(\mathbf{T}), \tag{6}$$

$$\mathcal{U}(\boldsymbol{\theta}, \boldsymbol{\omega}) = \{\mathbf{T} \geq 0 : \mathbf{T}\mathbf{1} = \boldsymbol{\theta},\; \mathbf{T}^\top \mathbf{1} = \boldsymbol{\omega}\}, \tag{7}$$

where $H(\mathbf{T}) = -\sum_{i,j} T_{ij} \log T_{ij}$ denotes the entropy term that encourages a smoother matching and $\eta > 0$ balances transport fidelity and smoothness. The plan $\mathbf{T}^\star$ is computed via differentiable Sinkhorn iterations. The marginals encode the expected mass per text token and per patch, enforcing globally consistent alignment of the attention distribution.

In practice, we set the patch prior to the uniform distribution over patches, i.e. $\boldsymbol{\omega} = \frac{1}{N}\mathbf{1}$. To prevent over-peaked text attention, we form the text prior by mixing similarity scores with a uniform term: for each text token $\mathbf{t}_i^p$ and pathway global visual token $\widetilde{\mathbf{f}}_p$, let $s_i = \cos(\mathbf{t}_i^p, \widetilde{\mathbf{f}}_p)$ and

$$\theta_i = \alpha_1 \frac{\exp(s_i)}{\sum_{k=1}^{L} \exp(s_k)} + (1 - \alpha_1)\frac{1}{L}, \tag{8}$$

where $\alpha_1 \in [0, 1]$ is a learnable blending coefficient that controls the mixture strength. Since $\mathbf{T}^\star$ is a coupling whose row sums equal $\boldsymbol{\theta}$, we row-normalize it to obtain conditional alignments on patches, i.e. $\overline{T}_{ij} = T^\star_{ij}/(\theta_i + \varepsilon)$, where ε is a small constant for numerical stability. To enable the attention mechanism to internalize the OT-induced distribution, we impose a distributional regularization via Jensen-Shannon (JS) divergence:

$$\mathcal{L}_{ot} = D_{\mathrm{JS}}(\mathbf{S} \parallel \overline{\mathbf{T}}) = \frac{1}{2L} \sum_{i=1}^{L} \Big[D_{\mathrm{KL}}(\mathbf{S}_i \parallel \mathbf{M}_i) + D_{\mathrm{KL}}(\overline{\mathbf{T}}_i \parallel \mathbf{M}_i) \Big], \tag{9}$$

where $\mathbf{M}_i = (\mathbf{S}_i + \overline{\mathbf{T}}_i)/2$. After obtaining the row-normalized OT plan $\overline{\mathbf{T}}$, we replace the local attention weights with globally regularized ones to aggregate visual evidence. The attended visual descriptor is computed as $\widetilde{\mathbf{V}} = \overline{\mathbf{T}}\mathbf{V}_v$, where each row corresponds to a text token enriched by globally matched visual cues. We then inject the attended visual evidence via a residual update, and then apply a pre-norm FFN:

$$\mathbf{z}_p = \mathbf{t}^p + \widetilde{\mathbf{V}}, \qquad \widetilde{\mathbf{t}}^p = \mathbf{z}_p + \mathrm{FFN}\big(\mathrm{LayerNorm}(\mathbf{z}_p)\big). \tag{10}$$

Given each input image $\mathbf{x}$, we compute branch-wise probabilities to align the attribute, object, and composition pathways with their corresponding primitives and compositions:

$$p\left(a_i \mid \mathbf{x}\right) = \frac{\exp(\widetilde{\mathbf{t}}_i^a \cdot \widetilde{\mathbf{f}}_a/\tau_c)}{\sum_{k=1}^{|\mathcal{A}|} \exp(\widetilde{\mathbf{t}}_k^a \cdot \widetilde{\mathbf{f}}_a/\tau_c)}, \quad p\left(o_j \mid \mathbf{x}\right) = \frac{\exp(\widetilde{\mathbf{t}}_j^o \cdot \widetilde{\mathbf{f}}_o/\tau_c)}{\sum_{k=1}^{|\mathcal{O}|} \exp(\widetilde{\mathbf{t}}_k^o \cdot \widetilde{\mathbf{f}}_o/\tau_c)}, \tag{11}$$

$$p\left(c_{ij} \mid \mathbf{x}\right) = \frac{\exp(\mathbf{t}_{ij}^c \cdot \mathbf{f}_c/\tau_c)}{\sum_{k=1}^{|\mathcal{C}_s|} \exp(\mathbf{t}_k^c \cdot \mathbf{f}_c/\tau_c)}, \tag{12}$$

where τ_c is the temperature parameter inherited from CLIP. Minimizing the negative log-likelihood over the training set $\mathcal{T}$ for the attribute, object, and composition pathways yields:

$$\mathcal{L}_{cls} = -\mathbb{E}_{(\mathbf{x}, c_{ij}) \sim \mathcal{T}} \left[\log p(a_i \mid \mathbf{x}) + \log p(o_j \mid \mathbf{x}) + \log p(c_{ij} \mid \mathbf{x}) \right]. \tag{13}$$

Finally, the overall BAD objective combines the classification losses with OT regularization applied to both the attribute and object pathways:

$$\mathcal{L}_{\text{BAD}} = \mathcal{L}_{cls} + \lambda_{ot}\big(\mathcal{L}^{a}_{ot} + \mathcal{L}^{o}_{ot}\big), \tag{14}$$

where λ_{ot} controls the regularization strength.

3.4 Prototype-driven Compositional Modulation

After BAD, the attribute, object, and composition branches are aligned with their respective primitives or compositions. Nonetheless, this three-branch design still depends on implicit associations and lacks explicit modeling of semantic interplay between primitives, hindering generalization to unseen compositions. To explicitly capture cross-primitive compositional semantics, PCM keeps all BAD-learned parameters frozen and maintains a learnable bank of attribute prototypes as robust semantic carriers. The selected prototype conditionally modulates the object features, making cross-primitive interactions explicit.

Reliable Prototype Selection. Formally, let $\mathbf{P} \in \mathbb{R}^{M\times D}$ denote the learnable global bank of M attribute prototypes. To systematically infuse these prototypes with fine-grained, instance-specific spatial context, we employ a lightweight Query-Transformer module designed to dynamically aggregate attribute-salient patch tokens, $\widetilde{\mathbf{F}}_a \in \mathbb{R}^{N\times D}$, retrieved from the BAD stage. This context-aware allocation is formalized as:

$$\mathbf{P}_t = \mathcal{F}_{\text{Q}}\big(\mathbf{P}, \widetilde{\mathbf{F}}_a\big), \tag{15}$$

where $\mathcal{F}_{\text{Q}}(\cdot,\cdot)$ operates as a cross-attention mechanism, establishing soft, deterministic correspondences between abstract prototype queries and localized visual attribute manifestations. This dynamic refinement enables the prototypes to capture both invariant attribute semantics and context-sensitive variations. The refined prototypes are then integrated with the original prototype bank:

$$\widetilde{\mathbf{P}} = (1-\alpha_2)\mathbf{P}_t + \alpha_2\mathbf{P}, \tag{16}$$

where $\alpha_2 \in [0,1]$ is a learnable blending coefficient. The dynamic prototypes $\mathbf{P}_t$ capture image-level attribute evidence from patch tokens and are subsequently fused with the global prototype bank to balance instance-adaptive refinement and global semantic stability.

To ensure that the global prototypes learn instance-invariant attribute semantics while maintaining diversity in the representation space, we introduce two complementary regularizers. Let $\overline{\mathbf{P}} \in \mathbb{R}^{M\times D}$ denote ℓ_2-normalized global prototypes. The similarity matrix $\mathbf{G} = \overline{\mathbf{P}}(\mathbf{t}^a)^\top$ induces two bidirectional assignment distributions: $\mathbf{A}_{v\to t}$ and $\mathbf{A}_{t\to v}$, obtained by applying the softmax function along the rows and columns of $\mathbf{G}$, respectively. To align the cross-modal semantics, we enforce consistency between these distributions via the Jensen-Shannon divergence. Simultaneously, to prevent prototype collapse, we impose the orthogonality constraint on $\overline{\mathbf{P}}$:

$$\mathcal{L}_{align} = D_{\text{JS}}(\mathbf{A}_{v\to t} \parallel \mathbf{A}^{\top}_{t\to v}), \quad \mathcal{L}_{orth} = \big\|\overline{\mathbf{P}}(\overline{\mathbf{P}})^\top - \mathbf{I}_M\big\|^2_{\text{F}}, \tag{17}$$

where $\mathbf{I}_M \in \mathbb{R}^{M\times M}$ represents the identity matrix.

To identify the most relevant attribute prototype for a given image, we formulate prototype selection as a differentiable sampling process over a categorical distribution. Given the global attribute token $\widetilde{\mathbf{f}}_a$ and the blended prototype bank $\widetilde{\mathbf{P}}$, we apply the continuous relaxation of the Gumbel-Max trick [9] to compute a sampling weight vector $\mathbf{w} \in \mathbb{R}^M$:

$$w_m = \frac{\exp((\pi_m + g_m)/\tau_g)}{\sum_{k=1}^{M} \exp((\pi_k + g_k)/\tau_g)}, \quad g_m \sim \text{Gumbel}(0, 1), \tag{18}$$

where $\pi_m = \langle \widetilde{\mathbf{P}}_{m,:}, \widetilde{\mathbf{f}}_{\text{a}} \rangle$ denotes the similarity logit for the m-th prototype, g_m represents independent and identically distributed noise samples, and τ_g is the temperature parameter, which is fixed to 0.1. The selected attribute prototype is then obtained via weighted aggregation as $\mathbf{p}^\star = \mathbf{w}^\top \widetilde{\mathbf{P}}$.

Compositional Modulation. Explicit compositional feature modeling requires moving beyond naive concatenation or addition operations, which fail to capture the conditional semantics of attributes applied to objects. The essence of human compositional imagination lies in the dynamic extrapolation of abstract attributes onto concrete visual entities to synthesize "virtual" representations. This cognitive mechanism is closely analogous to feature-wise linear modulation [26] within deep neural networks. Driven by this intuition, the compositional "imagination" mechanism is mathematically formalized as a conditional mapping function $\Phi : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^D$. Specifically, the targeted attribute prototype $\mathbf{p}^\star$ serves as a semantic prior to dynamically generate affine scaling and shifting factors, thereby reshaping the object representation $\widetilde{\mathbf{f}}_o$ within the feature manifold:

$$\widetilde{\mathbf{f}}_{ic} = \Phi(\widetilde{\mathbf{f}}_o \mid \mathbf{p}^\star) = \boldsymbol{\gamma}(\mathbf{p}^\star) \odot \widetilde{\mathbf{f}}_o + \boldsymbol{\beta}(\mathbf{p}^\star), \tag{19}$$

where $\odot$ denotes the Hadamard product, and $\boldsymbol{\gamma}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ are non-linear projections parameterized by separate MLPs. By explicitly conditioning the object features on the attribute prototype, this modulation mechanism achieves deep cross-primitive semantic coupling. To ensure proper alignment between the imagined representation $\widetilde{\mathbf{f}}_{ic}$ and its corresponding compositional concept, the probability distribution and the imaginative composition loss are computed as:

$$p_{ic}(c_{ij}|\mathbf{x}) = \frac{\exp(\mathbf{t}_{ij}^c \cdot \widetilde{\mathbf{f}}_{ic}/\tau_c)}{\sum_{k=1}^{|\mathcal{C}_s|} \exp(\mathbf{t}_k^c \cdot \widetilde{\mathbf{f}}_{ic}/\tau_c)}, \quad \mathcal{L}_{ic} = -\mathbb{E}_{(\mathbf{x},c_{ij})\sim\mathcal{T}} \log p_{ic}(c_{ij} \mid \mathbf{x}). \tag{20}$$

The overall training objective for PCM combines the imaginative composition loss with alignment and orthogonality regularization:

$$\mathcal{L}_{\text{PCM}} = \mathcal{L}_{ic} + \lambda_{orth}\mathcal{L}_{orth} + \lambda_{align}\mathcal{L}_{align}. \tag{21}$$

Inference. During testing, predictions from the BAD stage are combined with imaginative composition scores from PCM to form the final output. The predicted category is obtained by integrating probabilities from all pathways:

$$\hat{c} = \underset{c_{ij}\in\mathcal{C}_t}{\arg\max} \quad \alpha_{ic} p_{ic}(c_{ij}|\mathbf{x}) + \alpha_c p(c_{ij}|\mathbf{x}) + p(a_i|\mathbf{x}) \cdot p(o_j|\mathbf{x}), \tag{22}$$

Table 1. Experimental results on UT-Zappos, MIT-States, and C-GQA across both *Closed-world* and *Open-world* settings. Top performances are highlighted in **bold**.

METHOD	UT-Zappos				MIT-States				C-GQA			
	Unseen↑	Seen↑	HM↑	AUC↑	Unseen↑	Seen↑	HM↑	AUC↑	Unseen↑	Seen↑	HM↑	AUC↑
Closed-world Results												
CLIP [29] (ICML'21)	49.1	15.8	15.6	5.0	30.2	46.0	26.1	11.0	25.0	7.5	8.6	1.4
CoOp [40] (IJCV'22)	49.3	52.1	34.6	18.8	47.6	34.4	29.8	13.5	26.8	20.5	17.1	4.4
CSP [25] (ICLR'23)	66.2	64.2	46.6	33.0	49.9	46.6	36.3	19.4	26.8	28.8	20.3	6.2
DFSP [19] (CVPR'23)	71.7	66.7	47.2	36.0	52.0	46.9	37.3	20.6	32.0	38.2	27.1	10.5
Troika [7] (CVPR'24)	73.8	66.8	54.6	41.7	53.0	49.0	39.3	22.1	35.7	41.0	29.4	12.4
PLID [2] (ECCV'24)	68.8	67.3	52.4	38.7	52.4	49.7	39.0	22.1	33.0	38.8	27.9	11.0
MSCI [32] (IJCAI'25)	75.5	67.4	59.2	45.8	53.4	50.2	39.9	22.8	38.2	42.4	31.7	14.2
Logic [33] (CVPR'25)	74.9	69.6	57.8	45.8	53.9	50.8	40.5	23.4	39.4	44.4	33.3	15.3
CLUSPRO [28] (ICLR'25)	76.0	70.7	58.5	46.6	54.0	52.1	40.7	23.8	37.8	44.3	32.8	14.9
VP-CMJL [38] (ICCV'25)	76.3	71.9	58.5	47.9	52.6	51.8	40.4	23.3	40.2	46.0	34.9	16.3
Ours	76.0	71.9	**60.4**	**49.0**	53.9	52.7	**40.8**	**24.1**	42.7	46.5	**36.4**	**17.6**
Open-world Results												
CLIP [29] (ICML'21)	20.6	15.7	11.2	2.2	14.3	30.1	12.8	3.0	4.6	7.5	4.0	0.3
CoOp [40] (IJCV'22)	31.5	52.1	28.9	13.2	9.3	34.6	12.3	2.8	4.6	21.0	5.5	0.7
CSP [25] (ICLR'23)	44.1	64.1	38.9	22.7	15.7	46.3	17.4	5.7	5.2	28.7	6.9	1.2
DFSP [19] (CVPR'23)	60.0	66.8	44.0	30.3	18.5	47.5	19.3	6.8	7.2	38.3	10.4	2.4
Troika [7] (CVPR'24)	61.2	66.4	47.8	33.0	18.4	48.8	20.1	7.2	7.9	40.8	10.9	2.7
PLID [2] (ECCV'24)	55.5	67.6	46.6	30.8	18.7	49.1	20.4	7.3	7.5	39.1	10.6	2.5
MSCI [32] (IJCAI'25)	63.0	67.4	53.2	37.3	20.6	49.2	21.2	7.9	10.6	42.0	13.7	3.8
Logic [33] (CVPR'25)	63.7	69.6	50.8	36.2	21.4	50.7	22.4	8.7	9.3	43.7	12.6	3.4
CLUSPRO [28] (ICLR'25)	66.2	71.0	54.1	39.5	22.1	51.2	23.0	9.3	8.3	41.6	11.6	3.0
VP-CMJL [38] (ICCV'25)	66.6	71.9	54.5	41.4	19.9	51.8	22.0	8.3	11.5	46.0	15.5	4.6
Ours	68.2	72.5	**56.1**	**42.1**	22.7	52.5	**23.6**	**9.6**	12.0	46.4	**15.6**	**4.8**

where α_c and α_{ic} denote hyper-parameters that balance contributions from different compositional pathways.

4 Experiment

4.1 Experimental Setup

Datasets. We evaluate our approach on four established compositional zero-shot learning benchmarks: UT-Zappos [37], MIT-States [8], C-GQA [24], and Clothing16K [39]. These datasets exhibit diverse scales and semantic complexity. MIT-States encompasses 53,753 natural images annotated with 115 attributes and 245 objects. UT-Zappos contains 50,025 fine-grained shoe images with 16 attributes, 12 objects, and 116 valid compositions. C-GQA provides the broadest coverage, featuring 39,298 images, 453 attributes, 870 objects, and over 9,500 compositional pairs. Clothing16K is a low-noise apparel dataset with 16,170 images covering 9 color attributes and 8 clothing objects. We adopt the standard data splits from prior studies [7,25] for fair comparison. A comprehensive description of the datasets can be found in the supplementary material.

Metrics. We adopt standard CZSL metrics to comprehensively assess model performance. Specifically, we report the best accuracy on seen compositions and unseen compositions, along with their harmonic mean (HM) to evaluate balanced recognition capability. Additionally, we compute the area under the seen-unseen accuracy curve (AUC) by sweeping a calibration threshold, which provides a holistic measure of overall compositional generalization.

Table 2. *Closed-world* and *Open-world* results on Clothing16K.

METHOD	*Closed-world*				*Open-world*			
	Unseen↑	Seen↑	HM↑	AUC↑	Unseen↑	Seen↑	HM↑	AUC↑
CGE [24] (CVPR'21)	97.4	98.0	84.2	89.2	69.7	98.5	68.3	62.0
Co-CGE [21] (TPAMI'22)	94.7	98.5	87.9	88.3	63.8	98.7	69.2	59.3
SCEN [17] (CVPR'22)	89.6	98.0	78.5	78.8	62.3	96.7	61.5	53.7
OADis [30] (CVPR'22)	94.2	97.7	86.1	88.4	58.6	98.0	63.2	53.4
ADE [5] (CVPR'23)	97.7	98.2	88.7	92.4	73.1	99.0	74.2	68.0
CLPS [13] (TMM'25)	98.9	99.2	91.7	96.2	76.1	99.2	76.1	71.5
ATIF [22] (TMM'25)	99.4	99.0	95.4	96.6	93.9	99.2	92.0	91.8
AIF [22] (TMM'25)	99.6	99.5	96.3	98.2	92.2	99.7	90.3	89.9
Ours	99.9	99.7	**98.6**	**99.2**	97.1	100.0	**95.5**	**96.4**

Implementation Details. All experiments are conducted on a single NVIDIA RTX A6000 GPU using the PyTorch framework. We adopt CLIP with a ViT-L/14 visual backbone and its corresponding text encoder. By default, we perform 200 iterations of the Sinkhorn algorithm [3] to solve the OT problem, with the entropic regularization parameter η set to 0.5. The hyperparameters α_c and α_{ic} are tuned based on validation performance. Additional implementation details and parameter configurations are provided in Appendix.

4.2 Comparison with State-of-the-Art Methods

We conduct extensive comparisons with existing CZSL methods under both *Closed-world* and *Open-world* settings on three standard benchmarks: UT-Zappos [37], MIT-States [8], and C-GQA [24]. The evaluated methods include CLIP [29], CoOp [40], CSP [25], DFSP [19], Troika [7], PLID [2], MSCI [32], Logic [33], CLUSPRO [28] and VP-CMJL [38]. Comprehensive results are summarized in Table 1. Furthermore, we evaluate PCI against established baselines on Clothing16K [39], with comparative results presented in Table 2. The benchmarked methods include CGE [24], Co-CGE [21], SCEN [17], OADis [30], ADE [5], CLPS [13], ATIF [22], and AIF [22].

Closed-world. Across all four benchmarks, PCI delivers the best AUC. On Clothing16K, PCI achieves an AUC of **99.2**, surpassing previous methods by a significant margin. On UT-Zappos, it achieves an AUC of **49.0**, demonstrating a substantial absolute gain of **+1.1%** over the strongest prior art. On MIT-States, despite substantial attribute ambiguity and annotation noise, PCI still sets a new SOTA AUC. On the large-scale C-GQA, PCI attains **17.6** AUC, establishing a new state of the art. Beyond AUC, the accompanying Seen/Unseen accuracies remain well balanced, indicating improved transfer to novel compositions without sacrificing seen performance.

Open-world. The *Open-world* setting is more challenging due to an enlarged candidate space, yet PCI remains ahead in AUC and HM on all datasets. On Clothing16K, PCI attains an AUC of **96.4**, representing a notable **4.6%** improvement over the previous state-of-the-art. For UT-Zappos, PCI reaches **42.1** AUC, outperforming the strongest competitor VP-CMJL [38] by **0.7%**, while also achieving the highest AUC on both MIT-States and C-GQA. Consistent

Table 3. Main ablation studies of key components.

BAD		PCM	UT-Zappos				MIT-States			
VE	SD		Unseen↑	Seen↑	HM↑	AUC↑	Unseen↑	Seen↑	HM↑	AUC↑
Closed-world										
✗	✗	✗	74.4	65.4	55.7	42.1	52.3	50.2	39.3	22.3
✔	✗	✗	73.2	68.2	57.6	44.2	52.0	51.4	39.9	22.7
✗	✔	✗	71.8	68.1	57.1	43.2	52.7	51.0	39.5	22.6
✔	✔	✗	74.6	72.1	58.3	47.2	52.1	53.3	40.1	23.4
✔	✔	✔	76.0	71.9	**60.4**	**49.0**	53.9	52.7	**40.8**	**24.1**
Open-world										
✗	✗	✗	61.9	68.4	53.2	36.3	21.4	46.3	21.4	7.9
✔	✗	✗	67.5	69.1	51.8	38.4	21.8	50.6	22.4	8.7
✗	✔	✗	64.6	69.3	51.1	37.3	22.0	49.7	22.1	8.6
✔	✔	✗	65.8	70.9	54.7	40.0	22.2	51.6	22.9	9.1
✔	✔	✔	68.2	72.5	**56.1**	**42.1**	22.7	52.5	**23.6**	**9.6**

improvements in HM and Seen/Unseen metrics further corroborate PCI's robustness under significant distribution shifts.

4.3 Ablation Study

Key Component Analysis. We ablate five configurations on UT-Zappos and MIT-States (Table 3): the 3-branch baseline, +VE, +SD, BAD (VE+SD), and the full model. Adding VE improves HM/AUC by injecting text-guided priors into visual tokens and strengthening local evidence aggregation. Introducing SD alone also yields sizable gains; its OT-based global mass allocation enforces competition across patches and produces globally consistent text–patch correspondences, alleviating ambiguous matching within each primitive. Combining the two (BAD) delivers the strongest alignment within the 3-branch stage and steadily raises both seen and unseen accuracy. Integrating PCM with BAD yields the most significant performance gains in the Unseen and HM metrics. These results suggest that PCM provides an explicit attribute-conditioned interaction mechanism, which improves generalization to unseen compositions and yields a more favorable seen–unseen trade-off.

Effects of SD. We evaluate SD via a progressive replacement of the alignment block with three variants: (i) Vanilla—replace the OT-based alignment with standard multi-head attention while keeping all other components fixed; (ii) SD (w/o $\mathcal{L}_{ot}$)—restore the entropic OT solver to obtain a transport plan but remove the distributional regularizer; and (iii) SD (w/ $\mathcal{L}_{ot}$)—the full model. Results are reported in Table 4. With OT enabled, Sinkhorn iterations impose text–patch marginal constraints and induce competition for limited mass across patches, yielding globally consistent correspondences and mitigating cross-primitive entanglement. Adding $\mathcal{L}_{ot}$ further aligns the model's native attention $\mathbf{S}$ with the OT plan $\overline{\mathbf{T}}$, prompting the network to internalize the globally regularized distribution and improving image–text calibration.

Effects of Compositional Modulation in PCM. After obtaining reliable attribute prototype priors in PCM, various methods can be employed to generate compositional "imagined" features. To validate the necessity of our Compositional Modulation (CM) module, we replace it with several alternative feature

Table 4. Effects of SD. See § 4.3 for more details.

METHOD	UT-Zappos				MIT-States			
	Unseen↑	Seen↑	HM↑	AUC↑	Unseen↑	Seen↑	HM↑	AUC↑
Closed-world								
Vanilla	74.6	70.3	57.5	45.8	52.3	51.5	39.6	22.8
SD (w/o $\mathcal{L}_{ot}$)	74.2	71.0	58.0	46.7	52.2	52.8	39.8	23.1
SD (w $\mathcal{L}_{ot}$)	74.6	72.1	**58.3**	**47.2**	52.1	53.3	**40.1**	**23.4**
Open-world								
Vanilla	67.6	69.0	52.3	38.4	22.1	47.9	21.9	8.4
SD (w/o $\mathcal{L}_{ot}$)	65.1	70.8	**54.9**	39.7	21.8	50.6	22.3	8.8
SD (w $\mathcal{L}_{ot}$)	65.8	70.9	54.7	**40.0**	22.2	51.6	**22.9**	**9.1**

Table 5. Effects of Compositional Modulation in PCM. See § 4.3 for more details.

METHOD	UT-Zappos				MIT-States			
	Unseen↑	Seen↑	HM↑	AUC↑	Unseen↑	Seen↑	HM↑	AUC↑
Closed-world								
Add	75.4	70.5	58.7	47.4	52.1	52.9	40.1	23.5
MLP	76.3	70.7	57.5	47.3	52.8	53.6	40.4	23.8
Bilinear	75.6	71.8	58.5	47.8	52.4	53.0	40.2	23.4
CM	76.0	71.9	**60.4**	**49.0**	53.9	52.7	**40.8**	**24.1**
Open-world								
Add	67.0	71.5	55.7	41.2	22.6	51.6	23.1	9.3
MLP	67.3	70.4	55.3	41.0	22.7	51.9	23.3	9.4
Bilinear	67.0	70.5	55.3	40.8	22.8	51.2	23.3	9.3
CM	68.2	72.5	**56.1**	**42.1**	22.7	52.5	**23.6**	**9.6**

generation strategies, including direct addition of attribute prototypes and object features (Add), concatenation followed by MLP projection (MLP), and Bilinear fusion. Experimental results presented in Table 5 demonstrate that our CM module achieves superior performance across both datasets. This indicates that feature modulation better aligns with human cognition in imagining virtual compositional features, where attribute priors exert conditional influence on the current object to generate meaningful compositions.

5 Qualitative Results

Case Studies. In Fig. 3, we visualize qualitative results of both seen and unseen classes for comprehensive case analysis. We report the predictions of the BAD model along with the Top-1 predictions of the full PCI model. It is observed that while competitive results can be achieved through sufficient visual–language alignment in BAD, the model tends to produce ambiguous outcomes in cases such as "`white train`," primarily due to its independent primitive-level decision-making, which fails to capture compositional relationships across primitives. In contrast, PCI addresses this limitation by explicitly modeling primitive interactions and leveraging synthesized virtual compositional features, thereby enhancing distinctions between concepts like "`white train`" and "`yellow train`." For a thorough analysis, we also report some failure cases of PCI. Although the predictions of PCI do not match the provided labels, they remain interpretable and reasonably describe the image.

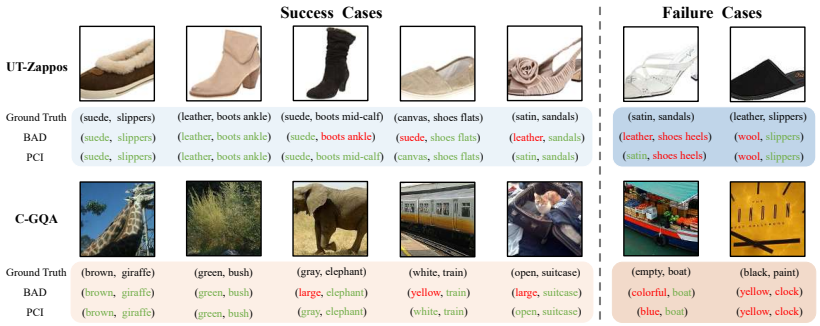

Fig. 3. We show the top-1 predictions of PCI together with the BAD results. The first five columns present successful PCI predictions, while the last two columns highlight representative failures. Green denotes correct predictions, and red marks incorrect ones.

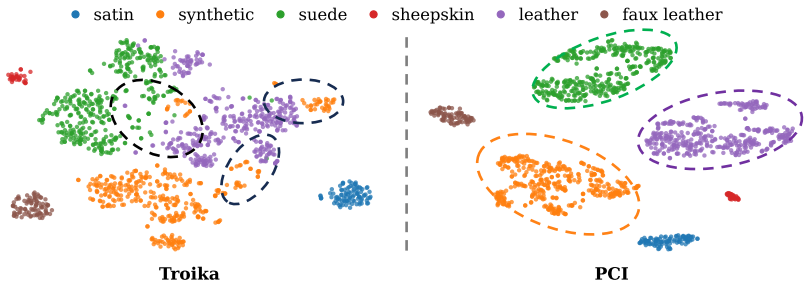

Fig. 4. t-SNE visualization of dynamic-attribute text features learned on UT-Zappos: Troika [7] (left) vs. PCI (right).

Visualization of Attribute Feature Distributions. Both Troika [7] and PCI aim to enrich text representations for better adaptation to diverse visual modalities in the real world. The key distinction is that the SD module in PCI leverages optimal transport to enforce globally consistent alignment. We visualize the dynamic attribute text features to illustrate the learned representations. As shown in Fig. 4, the globally consistent alignment induced by optimal transport enables PCI to produce more compact intra-class clusters and more separated inter-class distributions. Compared with Troika, even for visually or semantically similar concepts such as "`synthetic`", "`suede`", and "`leather`", PCI constructs a well-organized embedding space, demonstrating its superior alignment and disentanglement capabilities.

Contribution of PCM to Compositional Recognition. We visualize the contribution of α_{ic} in recognizing compositional concepts. Specifically, when α_{ic} approaches 0, the model predominantly relies on scores generated by BAD for prediction. Increasing α_{ic} assigns more weight to the PCM-based imagination scores. As shown in Fig. 5, AUC and HM peak when $\alpha_{ic} \in [0.2, 0.6]$, suggest-

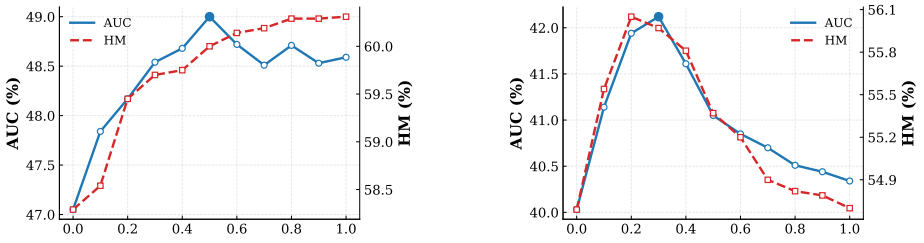

Fig. 5. Effect of PCM on compositional recognition for UT-Zappos. Left: *Closed-world* results; right: *Open-world* results.

ing that PCM complements BAD by modeling cross-primitive semantics and calibrating its predictions, thereby improving generalization.

6 Conclusion

We introduce Prototype-Conditioned Imagination, a two-stage framework for CZSL. PCI tackles challenges: (i) ambiguity from visually and semantically similar concepts, and (ii) the inability of models trained on isolated primitives to capture compositional pattern matching. In Stage-I, we employ bidirectional cross-modal attention to enrich visual representations and enforce consistent disentanglement. In Stage-II, operating on the disentangled and aligned space, we mimic human concept formation: attribute prototypes encode concepts and drive cross-primitive interactions to synthesize imagined compositional features.

Acknowledgements

This work was supported partly by the National Natural Science Foundation of China (NSFC) under Grants No. U25A20444 and No. 62371235, and partly by the Key Research and Development Plan of Jiangsu Province under Grant BE2023008-2.

References

1. Anwaar, M.U., Pan, Z., Kleinsteuber, M.: On leveraging variational graph embeddings for open world compositional zero-shot learning. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4645–4654 (2022)
2. Bao, W., Chen, L., Huang, H., Kong, Y.: Prompting language-informed distribution for compositional zero-shot learning. In: European Conference on Computer Vision. pp. 107–123. Springer (2024)
3. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
4. Haeusser, P., Mordvintsev, A., Cremers, D.: Learning by association–a versatile semi-supervised training method for neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 89–98 (2017)

5. Hao, S., Han, K., Wong, K.Y.K.: Learning attention as disentangler for compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15315–15324 (2023)
6. Hu, X., Wang, Z.: Leveraging sub-class discimination for compositional zero-shot learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 890–898 (2023)
7. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24005–24014 (June 2024)
8. Isola, P., Lim, J.J., Adelson, E.H.: Discovering states and transformations in image collections. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1383–1391 (2015)
9. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: International Conference on Learning Representations (2017)
10. Jiang, C., Wang, S., Long, Y., Li, Z., Zhang, H., Shao, L.: Imaginary-connected embedding in complex space for unseen attribute-object discrimination. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
11. Jiang, C., Zhang, H.: Revealing the proximate long-tail distribution in compositional zero-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2498–2506 (2024)
12. Jiang, C., Zhang, H.: Revealing the proximate long-tail distribution in compositional zero-shot learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2498–2506 (2024)
13. Jiang, H., Chen, C., Yang, X., Xu, C.: Compact latent primitive space learning for compositional zero-shot learning. IEEE Transactions on Multimedia (2025)
14. Kantorovich, L.V.: On the translocation of masses. Journal of mathematical sciences **133**(4) (2006)
15. Karthik, S., Mancini, M., Akata, Z.: Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9336–9345 (2022)
16. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
17. Li, X., Yang, X., Wei, K., Deng, C., Yang, M.: Siamese contrastive embedding network for compositional zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9326–9335 (2022)
18. Li, Y.L., Xu, Y., Mao, X., Lu, C.: Symmetry and group in attribute-object compositions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11316–11325 (2020)
19. Lu, X., Guo, S., Liu, Z., Guo, J.: Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23560–23569 (2023)
20. Mancini, M., Naeem, M.F., Xian, Y., Akata, Z.: Open world compositional zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5222–5230 (2021)
21. 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 **46**(3), 1545–1560 (2022)
22. Min, L., Fan, Z., Wang, S., Dou, F., Li, X., Wang, B.: Adaptive fusion learning for compositional zero-shot recognition. IEEE Transactions on Multimedia (2024)

23. Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: Composition with context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1792–1801 (2017)
24. Naeem, M.F., Xian, Y., Tombari, F., Akata, Z.: Learning graph embeddings for compositional zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 953–962 (2021)
25. Nayak, N.V., Yu, P., Bach, S.H.: Learning to compose soft prompts for compositional zero-shot learning. arXiv preprint arXiv:2204.03574 (2022)
26. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
27. Purushwalkam, S., Nickel, M., Gupta, A., Ranzato, M.: Task-driven modular networks for zero-shot compositional learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3593–3602 (2019)
28. Qu, H., Wei, J., Shu, X., Wang, W.: Learning clustering-based prototypes for compositional zero-shot learning. arXiv preprint arXiv:2502.06501 (2025)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
30. Saini, N., Pham, K., Shrivastava, A.: Disentangling visual embeddings for attributes and objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13658–13667 (2022)
31. Wang, Q., Liu, L., Jing, C., Chen, H., Liang, G., Wang, P., Shen, C.: Learning conditional attributes for compositional zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11197–11206 (2023)
32. Wang, Y., Xu, S., Zhu, X., Li, Y.: Msci: Addressing clip's inherent limitations for compositional zero-shot learning. arXiv preprint arXiv:2505.10289 (2025)
33. Wu, P., Lu, X., Hu, H., Xian, Y., Shen, J., Wang, W.: Logiczsl: Exploring logic-induced representation for compositional zero-shot learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30301–30311 (2025)
34. Xu, G., Kordjamshidi, P., Chai, J.: Prompting large pre-trained vision-language models for compositional concept learning. arXiv preprint arXiv:2211.05077 (2022)
35. Yan, X., Feng, S., Zhang, Y., Yang, J., Lin, Y., Fei, H.: Leveraging mllm embeddings and attribute smoothing for compositional zero-shot learning. arXiv preprint arXiv:2411.12584 (2024)
36. Yang, M., Deng, C., Yan, J., Liu, X., Tao, D.: Learning unseen concepts via hierarchical decomposition and composition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10248–10256 (2020)
37. Yu, A., Grauman, K.: Fine-grained visual comparisons with local learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 192–199 (2014)
38. Zhang, S., Yan, C., Liu, Y., Jing, C., Zhou, L., Wang, W.: Learning visual proxy for compositional zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2793–2802 (October 2025)
39. Zhang, T., Liang, K., Du, R., Sun, X., Ma, Z., Guo, J.: Learning invariant visual representations for compositional zero-shot learning. In: European conference on computer vision. pp. 339–355. Springer (2022)
40. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)