I2CD: An Invertible Causal Framework for Compositional Zero-Shot Learning via Disentangle-Compose-Disentangle

Zhaoquan Yuan[1,2], Zining Wang[1], Yuankang Pan[1], Ao Luo[1], Wei Li[1*], Xiao Wu[1], Changsheng Xu[3]

[1]School of Computing and Artificial Intelligence, Southwest Jiaotong University, China
[2]Manufacturing Industry Chain Collaboration Industrial Software Key Laboratory of Sichuan Province, Chengdu, China
[3]State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, CAS, China
{zqyuan, aoluo, liwei, wuxiaohk}@swjtu.edu.cn, {wangzn, pyk}@my.swjtu.edu.cn, csxu@nlpr.ia.ac.cn

Abstract

Compositional Zero-Shot Learning (CZSL) addresses the challenge of recognizing unseen attribute-object compositions in images, representing a fundamental challenge in artificial intelligence. Current approaches, which primarily focus on semantic alignment or distribution independence of primitives, have not achieved effective state-object decoupling and causal interventional invariance, limiting their performance on unseen compositions. To tackle this challenge, this study introduces I2CD (**I**nvertible **C**ausal framework via **D**isentangle-**C**ompose-**D**isentangle), a novel framework that integrates invertible neural networks with causal intervention techniques to achieve state-object disentanglement. The framework employs a disentangle-compose-disentangle mechanism for counterfactual generation within the disentangled representation space, ensuring that modifications to one primitive (attribute or object) maintain independence from the other, thus enabling robust causal disentanglement. Representational consistency is maintained through semantic alignment between initial disentangled representations and their recomposed-then-disentangled counterparts with corresponding textual concepts. Comprehensive evaluations on three benchmark datasets—MIT-States, UT-Zappos, and C-GQA—demonstrate the framework's effectiveness in achieving both disentanglement and compositional generalization in CZSL tasks.

Introduction

Compositional learning and reasoning is one of the key building blocks for human intelligence but remains largely absent in current artificial intelligence systems (Lake et al. 2017), and this lack of compositional generalization hinders machines from effectively handling novel combinations of learned concepts. Inspired by this human cognitive capabilities, Compositional Zero-Shot Learning (CZSL) (Misra, Gupta, and Hebert 2017) aims to enable machines to recognize novel combinations of states and objects by learning from previously observed compositions. In this framework, each composition comprises two components (i.e., a state and an object), with distinct sets of compositions used for training and testing. Despite significant progress, robustly generalizing to unseen compositions remains a challenging problem for AI systems (Naeem et al. 2021; Li et al. 2023; Kim et al. 2023; Huang et al. 2024).

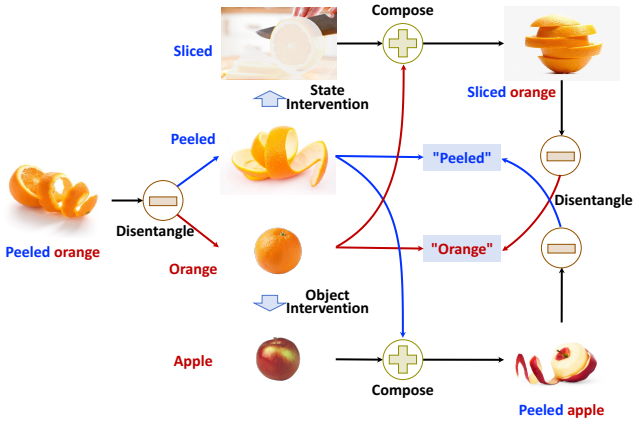

Figure 1: The illustration of our motivation.

CZSL presents several key challenges: 1) Entanglement: Semantic primitives are inherently entangled in the original pixel space, making it challenging to learn distinct representations for states and objects. These spurious correlations between features impair the model's ability to generalize to unseen compositions in zero-shot scenarios. 2) Contextuality: The visual manifestation of states can vary significantly across different objects, leading to context-dependent representations. This study primarily addresses the disentanglement challenge.

Existing approaches predominantly employ dual-branch architectures to extract distinct features for states and objects (Naeem et al. 2021; Ruis, Burghouts, and Bucur 2021; Qu et al. 2025). Some studies enhance attribute-object disentanglement through various techniques, including attention-based mechanisms (Hao, Han, and Wong 2023) and contrastive learning strategies (Li et al. 2022a; Hu and Wang 2023). Recent research has introduced causal perspectives to CZSL by conceptualizing the task as a causal intervention problem (Atzmon et al. 2020; Yang et al. 2023), where specific factors (attributes or objects) are modified while keeping others constant.

Despite these advancements, existing methods focus primarily on semantic alignment or distribution independence

*Corresponding author: Wei Li.

of primitives, failing to achieve true state-object decoupling and causal intervention invariance (Suter et al. 2019; Reddy, L, and Balasubramanian 2022). Consequently, these approaches struggle with unseen compositions, as novel combinations inherently represent interventions between primitives.

To overcome these limitations, we argue that semantic invariance must be maintained for non-intervened primitive representations within disentangled latent spaces. When performing an intervention on one primitive representation (such as state) and combining it with another unmodified primitive (such as object) to generate a new image, the subsequent disentanglement process should preserve the semantic integrity of the unmodified representation. The motivation is illustrated in Figure 1. Moreover, since modifications to primitives in the disentangled semantic space necessitate corresponding changes in the pixel space, the learning of disentangled representations requires a bijective mapping function, implemented through invertible neural networks.

Motivated by above considerations, we propose I2CD (**I**nvertible **C**ausal framework via **D**isentangle-**C**ompose-**D**isentangle), a novel framework for compositional zero-shot learning, as illustrated in Figure 2. Our I2CD framework builds upon the pre-trained CLIP (Radford et al. 2021) vision-language model, incorporating parameter-efficient adapters specifically designed for CZSL task. The framework employs an invertible neural network to disentangle images into their constituent state and object representations. To establish causal intervention invariance, we implement controlled interventions by substituting primitive representations of target images (e.g., state $\mathbf{h}_s^b$) with corresponding representations from reference images (e.g., state $\mathbf{h}_s^a$). These systematic substitutions simulate causal interventions within the disentangled representation space. Subsequently, the modified primitive representations are composed with the preserved representations (e.g., object $\mathbf{h}_o^b$) to synthesize novel images, which undergo a secondary disentanglement process to extract their respective state and object representations. To maintain representational consistency, we enforce semantic alignment between the initial disentangled representations and their recomposed-then-disentangled counterparts with their corresponding textual concepts. Also, the compositional representation semantic alignment is conducted in I2CD.

The key contributions of our work can be summarized as follows:

- We propose a novel invertible causal framework for compositional zero-shot learning that focuses on learning causally invariant disentangled representations of primitives through a disentangle-compose-disentangle mechanism.
- To simulate causal intervention and maintain semantic invariance for disentangled primitives, we substitute primitive features from reference images and synthesize counterfactual images. Both the primary and secondary disentangled representations are aligned with textual concepts.
- We conduct comprehensive experiments on three widely used CZSL benchmarks—MIT-States, UT-Zappos, and C-GQA—to validate the effectiveness and robustness of our I2CD framework compared to existing approaches.

Related Work

Compositional Zero-Shot Learning

Compositional Zero-Shot Learning (CZSL) has significantly evolved through diverse methodologies aiming to enhance the generalization to unseen attribute-object compositions. Existing approaches are generally fall into two categories: **1) Composed methods:** capture dependence of states and object, and focuse on learning compositional features. Early transformation-based methods employ contextual regularization, symmetry, and group theory to ensure consistent attribute-object relationships (Misra, Gupta, and Hebert 2017; Nagarajan and Grauman 2018; Li et al. 2022b; Purushwalkam et al. 2019). Graph-based approaches explicitly capture dependencies among attributes, objects, and their compositions using structured propagation mechanisms (Naeem et al. 2021; Mancini et al. 2022; Jiang et al. 2024b). Attention mechanisms contribute significantly by refining compositional feature interactions (Khan et al. 2023; Li et al. 2023; Kim et al. 2023). Conditional modeling methods treat attributes as dependent variables conditioned on recognized objects or visual contexts, significantly enhancing attribute adaptation in novel contexts (Wang et al. 2023b; Zhang et al. 2024). Prompt-based methods leverage large-scale Vision-Language Models (VLMs) for efficient CZSL, including compositional soft prompting (Nayak, Yu, and Bach 2023; Wang et al. 2023a), and Context-driven prompting (Lu et al. 2023; Li et al. 2024; Bao et al. 2024; Zheng, Zhu, and Nevatia 2024; Huang et al. 2024). **2) Disentangled methods:** emphasize attribute-object disentanglement to mitigate spurious correlations and improve generalization. To achieve disentanglement, attributes and objects are modeled conditionally independent (Atzmon et al. 2020; Ruis, Burghouts, and Bucur 2021). Recent advancements include attention-based disentanglers (Hao, Han, and Wong 2023), prompt-guided fusion modules (Lu et al. 2023), and complex embedding spaces modeling dynamic interactions via imaginary-real components (Jiang et al. 2024a).

While these methods improve disentanglement, they generally do not leverage counterfactual generation to enforce invariance under interventions. In contrast, our approach integrates invertible factorization with explicit causal intervention objectives, using counterfactuals to jointly guarantee semantic alignment and robust invariance—addressing a key limitation of prior causal CZSL frameworks.

Causal Disentanglement

Causal disentanglement represents a critical objective at the intersection of representation learning and causal inference. This approach focuses on developing representations where individual latent factors correspond to independent and interpretable causal mechanisms, thereby facilitating robust generalization, modular reasoning, and systematic intervention. The Independent Causal Mechanisms (ICM) principle (Schölkopf et al. 2021) and advances in

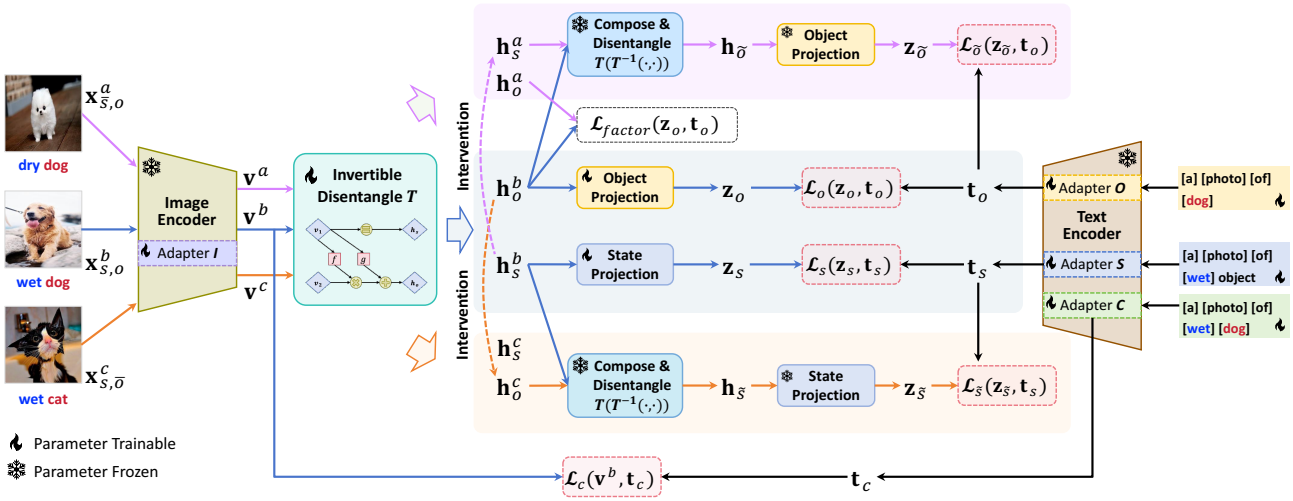

Figure 2: Overview of the proposed I2CD framework for compositional zero-shot learning.

causal meta-learning (Bengio et al. 2020) emphasize the significance of modularity and independence among generative factors. These principles have led to the development of various techniques for learning causally structured representations adaptable across domains and interventions, including causal effect estimation (Xu et al. 2023) and Interventional Robustness Score (IRS) (Suter et al. 2019). This study extends these principles by incorporating causal interventional invariance into Compositional Zero-Shot Learning (CZSL), ensuring that modifications to one semantic component remain independent of others, thus enabling robust and interpretable generalization to novel compositions.

The Proposed I2CD Framework

Problem Statement

In compositional zero-shot learning (CZSL), the objective is to recognize unseen compositions of known states and objects from images. Formally, let $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$ be the set of states and $\mathcal{O} = \{o_1, o_2, \ldots, o_{|\mathcal{O}|}\}$ be the set of objects. The compositional label space $\mathcal{Y}$ is defined as the Cartesian product of these two sets, i.e., $\mathcal{Y} = \{(s_i, o_j)|s_i \in \mathcal{S}, o_j \in \mathcal{O}\}$. The set of seen compositions $\mathcal{Y}_{\text{seen}} \subseteq \mathcal{Y}$ and unseen compositions $\mathcal{Y}_{\text{unseen}} \subseteq \mathcal{Y}$ are mutually exclusive and collectively exhaustive, where $\mathcal{Y}_{\text{seen}} \cap \mathcal{Y}_{\text{unseen}} = \emptyset$ and $\mathcal{Y}_{\text{seen}} \cup \mathcal{Y}_{\text{unseen}} = \mathcal{Y}$.

Given a training set $\mathcal{X}_{\text{train}} = \{(\mathbf{x}, \mathbf{y})|\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}_{\text{seen}}\}$, where $\mathcal{X}$ represents the image space, our objective is to develop a model that generalizes to unseen compositions $\mathcal{Y}_{\text{unseen}}$. Each training sample consists of an image paired with its corresponding compositional label, comprising a state and object pair. During evaluation, the model's performance is assessed on both seen and unseen compositions using the test set $\mathcal{X}_{\text{test}} = \{(\mathbf{x}, \mathbf{y})|\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}_{\text{test}}\}$.

In the generalized Compositional Zero-Shot Learning (CZSL) setting (Purushwalkam et al. 2019), the test set encompasses both seen and unseen compositions, defined as $\mathcal{Y}_{\text{test}} = \mathcal{Y}_{\text{seen}} \cup \mathcal{Y}'_{\text{unseen}}$. The model must predict labels for images containing both seen and unseen compositional combinations. CZSL evaluation primarily occurs in two scenarios: the closed-world and open-world settings. In the closed-world setting, $\mathcal{Y}'_{\text{unseen}}$ represents a predefined subset of $\mathcal{Y}_{\text{unseen}}$, formally expressed as $\mathcal{Y}'_{\text{unseen}} \in \mathcal{Y}_{\text{unseen}}$. The open-world setting (Mancini et al. 2021) presents a more challenging scenario where the target set incorporates all possible permutations of state-object compositions, expressed as $\mathcal{Y}_{\text{test}} = \mathcal{Y}$.

Overall Framework

We propose a novel framework for CZSL, termed I2CD (**I**nvertible **C**ausal framework via **D**isentangle-**C**ompose-**D**isentangle), as illustrated in Figure 2. The framework comprises three main components: 1) Feature embedding and compositional alignment; 2) Primitive semantic alignment; and 3) Causal interventional invariance.

Feature Embedding and Compositional Alignment

The feature embedding architecture in our I2CD is built on the pre-trained CLIP model (Radford et al. 2021). To optimize both computational efficiency and compositional flexibility, we implement an adapter-based fine-tuning approach for the image and text encoders (Nayak, Yu, and Bach 2023; Huang et al. 2024). The image encoder E_V employs the Vision Transformer (ViT) architecture (Dosovitskiy et al. 2021), incorporating lightweight visual adapters (Adapter *I*) within each transformer block. The text encoder E_T integrates three specialized intra-transformer adapters: Adapter *S* for state-specific features, Adapter *O* for object-specific features, and Adapter *C* for composition-specific features. These adapters, inspired by CAILA's concept-aware architecture (Zheng, Zhu, and Nevatia 2024), play a vital role in decomposing semantic components from the input label text.

To further improve alignment between visual and textual modalities, we employ a prompt tuning strategy. For each

branch (state, object, composition), an independent set of trainable prefix tokens is prepended to the input text, while the [state] and [object] tokens are shared across branches for semantic consistency. Specifically, the three prompt templates are defined as: $\mathbf{p}_s = [\text{prefix}_s, \text{[state]}]$, $\mathbf{p}_o = [\text{prefix}_o, \text{[object]}]$, and $\mathbf{p}_c = [\text{prefix}_c, \text{[state]}, \text{[object]}]$, where prefix_s, prefix_o, and prefix_c are independently learned for each semantic branch, with each prefix initialized to the phrase "*a photo of*". Passing these templates through the text encoder E_T with the corresponding adapters produces three specialized text embeddings:

$$\mathbf{t}_s = E_T(\mathbf{p}_s), \quad \mathbf{t}_o = E_T(\mathbf{p}_o), \quad \mathbf{t}_c = E_T(\mathbf{p}_c). \tag{1}$$

This mechanism ensures that attribute and object tokens are encoded consistently across the network while enabling branch-specific contextualization.

For a target image $\mathbf{x}^{\mathbf{b}}_{\mathbf{s},\mathbf{o}}$ with associated label $\mathbf{y} = (s_i, o_j)$, the ViT with visual adapters encodes the image into a compositional feature vector $\mathbf{v}^b = E_V(\mathbf{x}^b_{s,o})$, and the text encoder provides the corresponding compositional text embedding $\mathbf{t}_c$. The model is trained to align these multimodal embeddings using a similarity-based classification loss. Specifically, the probability of predicting label $\mathbf{y}$ for image $\mathbf{x}^b_{s,o}$ is computed as:

$$p(\mathbf{y}|\mathbf{x}^b_{s,o}) = \frac{\exp\big(\mathbf{v}^b \cdot \mathbf{t}_c/\tau_c\big)}{\sum_{k\in\mathcal{Y}_{\text{seen}}} \exp\big(\mathbf{v}^b \cdot \mathbf{t}^k_c/\tau_c\big)}, \tag{2}$$

where τ_c is a temperature parameter. The composition prediction loss is then defined as the cross-entropy over all training samples:

$$\mathcal{L}_c = -\frac{1}{|\mathcal{X}_{\text{train}}|} \sum_{(\mathbf{x}^{\mathbf{b}}_{\mathbf{s},\mathbf{o}}, \mathbf{y})\in\mathcal{X}_{\text{train}}} \log p(\mathbf{y}|\mathbf{x}^b_{s,o}). \tag{3}$$

Primitive Semantic Alignment

The core objective of this module is to disentangle the compositional visual representation $\mathbf{v}^b$ into two causally independent latent primitives, corresponding to the *state* and *object* components present in the image. This is achieved using an invertible network T based on a normalizing flow architecture (Esser, Rombach, and Ommer 2020), which establishes a bijective mapping between the compositional feature space and the disentangled latent space of equal dimension. Specifically, the forward mapping is defined as $(\mathbf{h}^b_s, \mathbf{h}^b_o) = T(\mathbf{v}^b)$, where $\mathbf{h}^b_s$ and $\mathbf{h}^b_o$ encode the state and object semantics, respectively.

To ensure that T learns a semantically meaningful and independent factorization, training leverages pairs of images that share one semantic factor while differing in the other. For instance, given an image $\mathbf{x}^b_{s,o}$ with state s and object o, we can identify another image $\mathbf{x}^a_{\bar{s},o}$ (sharing object o but differing in state) and another image $\mathbf{x}^c_{s,\bar{o}}$ (sharing state s but differing in object). These paired samples enable the model to attribute differences in the compositional representation to the correct underlying factor.

Crucially, the bijective nature of T guarantees not only that $\mathbf{h}^b_s$ and $\mathbf{h}^b_o$ can be independently manipulated, a property that underpins the subsequent causal intervention, but the original features can be reconstructed as $\mathbf{v}^b = T^{-1}(\mathbf{h}^b_s, \mathbf{h}^b_o)$. Independence between $\mathbf{h}^b_s$ and $\mathbf{h}^b_o$ is promoted by modeling their joint prior as factorized Gaussians: $p(\mathbf{h}^b_s, \mathbf{h}^b_o) = p(\mathbf{h}^b_s)\, p(\mathbf{h}^b_o)$, where each factor follows a standard normal distribution.

However, the independence prior alone is not sufficient to guarantee semantic alignment. We further constrain the latent factors to be invariant to changes in the other factor. For example, the object component $\mathbf{h}^b_o$ should remain consistent for pairs of images with the same object but different states. Let $(\mathbf{x}^a_{\bar{s},o}, \mathbf{x}^b_{s,o})$ be a pair sharing object o; their features $\mathbf{v}^a = E_V(\mathbf{x}^a_{\bar{s},o})$, $\mathbf{v}^b = E_V(\mathbf{x}^b_{s,o})$ are mapped by T to $(\mathbf{h}^a_s, \mathbf{h}^a_o)$ and $(\mathbf{h}^b_s, \mathbf{h}^b_o)$. We model the object factor as correlated Gaussian variables: $\mathbf{h}^b_o \sim \mathcal{N}\left(\mathbf{h}^b_o|\sigma_{ab}\,\mathbf{h}^a_o,\ (1-\sigma^2_{ab})I\right)$, where $\sigma_{ab} \in (0,1)$ denotes the expected correlation of the state factor. For the state factor in this pair, independence is enforced: $\mathbf{h}^a_s \sim \mathcal{N}(\mathbf{h}^a_s|0, I)$.

The training objective for T maximizes the likelihood of observed data under this factorized model using the change-of-variables formula for invertible flows. For any image pair sharing factor o, the joint probability is: $p(\mathbf{v}^b, \mathbf{v}^a|o) = p(\mathbf{v}^b)\, p(\mathbf{v}^a|\mathbf{v}^b, o)$. which can be expressed as:

$$|T'\left(\mathbf{v}^b\right)|\, p\left(T(\mathbf{v}^b)\right) \cdot |T'(\mathbf{v}^a)|\, p\left(T(\mathbf{v}^a)|T(\mathbf{v}^b), o\right), \tag{4}$$

where $|T'(\cdot)|$ denotes the Jacobian determinant of T. Inspired by the study (Esser, Rombach, and Ommer 2020), the corresponding negative log-likelihood loss for each pair is:

$$\begin{aligned}\ell(\mathbf{v}^a, \mathbf{v}^b|o) = &\sum_{k\in\{s,o\}} \left(\|\mathbf{h}^b_k\|^2 - \log|T'(\mathbf{v}^b)|\right) \\ &+ \left(\|\mathbf{h}^a_s\|^2 - \log|T'(\mathbf{v}^a)|\right) + \frac{\|\mathbf{h}^a_o - \sigma_{ab}\,\mathbf{h}^b_o\|^2}{1-\sigma^2_{ab}},\end{aligned} \tag{5}$$

where the first two terms enforce prior regularization and invertibility, and the final term aligns the shared factor in correlated pairs.

Aggregating across all shared-factor pairs yields the overall concept factorization loss:

$$\begin{aligned}\mathcal{L}_{\text{factor}} = &\ \mathbb{E}_{(\mathbf{x}^b_{s,o}, \mathbf{x}^c_{s,\bar{o}})\sim p(\cdot|s)}\, \ell\left(E_V(\mathbf{x}^b_{s,o}), E_V(\mathbf{x}^c_{s,\bar{o}})|s\right) \\ &+\ \mathbb{E}_{(\mathbf{x}^b_{s,o}, \mathbf{x}^a_{\bar{s},o})\sim p(\cdot|o)}\, \ell\left(E_V(\mathbf{x}^b_{s,o}), E_V(\mathbf{x}^a_{\bar{s},o})|o\right).\end{aligned} \tag{6}$$

While $\mathcal{L}_{\text{factor}}$ encourages independent factorization, we further promote semantic alignment by projecting each latent factor to the multimodal embedding space. Two separate MLPs, $\Pi_s(\cdot)$ for states and $\Pi_o(\cdot)$ for objects, map $\mathbf{h}^b_s$ and $\mathbf{h}^b_o$ to projected features $\mathbf{z}_s$ and $\mathbf{z}_o$, i.e., $\mathbf{z}_s = \Pi_s(\mathbf{h}^a_s)$, and $\mathbf{z}_o = \Pi_o(\mathbf{h}^a_o)$, ensuring dimension compatibility with CLIP text embeddings. Let $\mathbf{t}_s$ and $\mathbf{t}_o$ be the CLIP embeddings for state s and object o. We define the classification probabilities $p(s|\mathbf{x}^b_{s,o})$ and $p(o|\mathbf{x}^b_{s,o})$ respectively. The associated cross-entropy losses are:

$$\mathcal{L}_s = -\frac{1}{|\mathcal{X}_{\text{train}}|} \sum_{\mathbf{x}_{s,o}\in\mathcal{X}_{\text{train}}} \log p(s|\mathbf{x}_{s,o}), \tag{7}$$

$$\mathcal{L}_o = -\frac{1}{|\mathcal{X}_{\text{train}}|} \sum_{\mathbf{x}_{s,o} \in \mathcal{X}_{\text{train}}} \log p(o|\mathbf{x}_{s,o}). \tag{8}$$

Minimizing $\mathcal{L}_s$ and $\mathcal{L}_o$ ensures that the learned state and object factors are not only statistically independent, but also semantically aligned to the ground-truth state and object categories, thus enabling interpretable and robust compositional generalization.

Causal Interventional Invariance

To enhance the learning of causally disentangled representations derived from the invertible network T, we introduce a novel disentangle-compose-disentangle mechanism. To simulate a causal intervention, we substitute one primitive features from reference images ($\mathbf{x}^a_{\tilde{s},o}$ and $\mathbf{x}^c_{s,\tilde{o}}$), while keeping the other factor unchanged. The inverse mapping is then used to recombine these altered factors, yielding a counterfactual compositional feature. In our I2CD, we consider interventions on both the state and object components to enforce causal disentanglement.

For an image pair sharing object o but with different states, we construct a counterfactual by combining the state factor of one with the object factor of the other, and then recover the object factor by applying the invertible network after state intervention:

$$\mathbf{h}_{\tilde{o}} = T_o\big(T^{-1}(\mathbf{h}^b_o, \mathbf{h}^a_s)\big). \tag{9}$$

This is then projected to the object embedding space, $\mathbf{z}_{\tilde{o}} = \Pi_o(\mathbf{h}_{\tilde{o}})$, and used to compute the probability of the original object:

$$p\big(o|T^{-1}(\mathbf{h}^b_o, \mathbf{h}^a_s)\big) = \frac{\exp(\mathbf{z}_{\tilde{o}} \cdot \mathbf{t}_o/\tau_o)}{\sum_{k=1}^{|\mathcal{O}|} \exp(\mathbf{z}_{\tilde{o}} \cdot \mathbf{t}^k_o/\tau_o)}. \tag{10}$$

The object-invariance loss is then:

$$\mathcal{L}_{\tilde{o}} = -\log p\big(o|T^{-1}(\mathbf{h}^b_o, \mathbf{h}^a_s)\big). \tag{11}$$

Minimizing $\mathcal{L}_{\tilde{o}}$ ensures that object predictions remain robust to state interventions. Analogously, the state-invariance loss is:

$$\mathcal{L}_{\tilde{s}} = -\log p\big(s|T^{-1}(\mathbf{h}^b_s, \mathbf{h}^c_o)\big). \tag{12}$$

The overall intervention loss is defined as the sum of both invariance losses:

$$\mathcal{L}_{\text{intervention}} = \mathcal{L}_{\tilde{o}} + \mathcal{L}_{\tilde{s}}. \tag{13}$$

By optimizing $\mathcal{L}_{\text{intervention}}$, the model is explicitly encouraged to satisfy the principle of interventional invariance: that is, changing one component in the latent space via a simulated intervention should not affect the other component's semantic prediction. This enforces a strong form of causal disentanglement in the learned representations and is essential for reliable compositional generalization in CZSL.

Training Objective and Inference

Our I2CD framework optimizes three key objectives simultaneously: discriminative recognition, effective disentanglement, and causal invariance. The total loss function is defined as:

$$\mathcal{L} = \mathcal{L}_c + \alpha\, \mathcal{L}_{\text{dis}} + \beta\, \mathcal{L}_{\text{interv}}, \tag{14}$$

where $\mathcal{L}_c$ is the composition classification loss, $\mathcal{L}_{\text{dis}}$ is the disentanglement loss, $\mathcal{L}_{\text{interv}}$ is the causal intervention loss, and α, β are hyperparameters that control the relative weight of each objective.

The disentanglement loss further decomposes as:

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{factor}} + \mathcal{L}_s + \mathcal{L}_o, \tag{15}$$

where $\mathcal{L}_{\text{factor}}$ enforces that the compositional feature $\mathbf{v}^b$ is effectively separated into independent state and object factors $(\mathbf{h}^b_s, \mathbf{h}^b_o)$, and $\mathcal{L}_s$, $\mathcal{L}_o$ align the visual state and object features with their respective text embeddings in the shared multimodal space.

The intervention loss is defined as:

$$\mathcal{L}_{\text{interv}} = \mathcal{L}_{\tilde{s}} + \mathcal{L}_{\tilde{o}}, \tag{16}$$

where $\mathcal{L}_{\tilde{s}}$ encourages invariance of the state prediction under object intervention, and $\mathcal{L}_{\tilde{o}}$ encourages invariance of the object prediction under state intervention.

During inference, the final prediction for a given input $\mathbf{x}$ is made by integrating the compositional probability with the product of state and object probabilities. For a candidate label $\mathbf{y} = (s_i, o_j)$, the final composition score is computed as:

$$\tilde{p}(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}) + p(s_i|\mathbf{x}) \cdot p(o_j|\mathbf{x}), \tag{17}$$

where $p(\mathbf{y}|\mathbf{x})$ is the direct compositional classification probability, and $p(s_i|\mathbf{x})$, $p(o_j|\mathbf{x})$ are the probabilities assigned to state s_i and object o_j by the disentangled branches, respectively. The final predicted label $\hat{\mathbf{y}}$ is then given by:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}_{\text{test}}} \left(\tilde{p}(\mathbf{y}|\mathbf{x})\right), \tag{18}$$

where $\mathcal{Y}_{\text{test}}$ denotes the set of all candidate compositions in the test set.

Experiments

Datasets

We conduct experiments on three notable CZSL benchmarks: MIT-States (Isola, Lim, and Adelson 2015), UT-Zappos (Yu and Grauman 2014), and C-GQA (Naeem et al. 2021). MIT-States is designed to study state transformations on visual objects, containing a large number of natural images sourced from the Internet via search engines. It encompasses 115 states and 245 objects. UT-Zappos is a dataset focusing on various types of shoes with fine-grained visual attributes, comprising 16 states and 12 objects. C-GQA, a split built on the Stanford GQA dataset originally proposed for Visual Question Answering, is extensive with 413 states and 674 objects. Detailed statistics and splits are summarized in the appendix.

Evaluation Protocols and Metrics

We evaluate our models using established CZSL protocols (Naeem et al. 2021; Nayak, Yu, and Bach 2023), which involve both closed-world and open-world settings. In the closed-world scenario, the unseen classes are predefined and limited, while in the open-world setting, the model considers all possible state-object compositions, significantly expanding the search space. Specifically, in C-GQA, the number of

	Closed-World Setting	MIT-States				UT-Zappos				C-GQA			
		S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
w/o CLIP	CGE (Naeem et al. 2021)	32.8	28.0	21.4	6.5	64.5	71.5	**60.5**	33.5	31.4	14.0	14.5	3.6
	Co-CGE (Mancini et al. 2022)	32.1	28.3	20.0	6.6	62.3	66.3	48.1	33.9	33.3	14.9	14.4	4.1
	SCEN (Li et al. 2022a)	29.9	25.2	18.4	5.3	63.5	63.1	47.8	32.0	28.9	25.4	17.5	5.5
	OADis (Saini, Pham, and Shrivastava 2022)	31.1	25.6	18.9	5.9	59.5	65.5	44.4	30.0	-	-	-	-
	CANet (Wang et al. 2023b)	29.0	26.2	17.9	5.4	61.0	66.3	47.3	33.1	30.0	13.2	14.5	3.3
	CoT (Kim et al. 2023)	34.8	31.5	23.2	7.8	-	-	-	-	34.0	18.8	17.5	5.1
	ADE (Hao, Han, and Wong 2023)	-	-	-	-	63.0	64.3	51.1	35.1	35.0	17.7	18.0	5.2
w CLIP	CLIP (Radford et al. 2021)	30.2	46.0	26.1	11.0	15.8	49.1	15.6	5.0	7.5	25.0	8.6	1.4
	CoOp (Zhou et al. 2022)	34.4	47.6	29.8	13.5	52.1	49.3	34.6	18.8	20.5	26.8	17.1	4.4
	CSP (Nayak, Yu, and Bach 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
	HPL (Wang et al. 2023a)	47.5	50.6	37.3	20.2	63.0	68.8	48.2	35.0	30.8	28.4	22.4	7.2
	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
	ProLT (Jiang and Zhang 2024)	49.1	51.0	38.2	21.1	66.0	70.1	49.4	36.1	39.5	32.9	27.7	11.0
	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
	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
	CAILA (Zheng, Zhu, and Nevatia 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
	I2CD (Ours)	50.6	**54.1**	**40.0**	23.3	**69.1**	**77.1**	57.9	**46.4**	**44.1**	**41.2**	**35.0**	**16.3**

Table 1: Performance comparison of various CZSL models on three datasets under the closed-world setting.

	Open-World Setting	MIT-States				UT-Zappos				C-GQA			
		S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
w/o CLIP	CGE (Naeem et al. 2021)	32.4	5.1	6.0	1.0	61.7	47.7	39.0	23.1	32.7	1.8	2.9	0.47
	Co-CGE (Mancini et al. 2022)	30.3	11.2	10.7	2.3	61.2	45.8	40.8	23.3	32.1	3.0	4.8	0.78
	KG-SP (Karthik, Mancini, and Akata 2022)	28.4	7.5	7.4	1.3	61.8	52.1	42.3	26.5	31.5	2.9	4.7	0.78
	DRANet (Li et al. 2023)	29.8	7.8	7.9	1.5	65.1	54.3	44.0	28.8	31.3	3.9	6.0	1.05
	ADE (Hao, Han, and Wong 2023)	-	-	-	-	62.4	50.7	44.8	27.1	35.1	4.8	7.6	1.42
w CLIP	CLIP (Radford et al. 2021)	30.1	14.3	12.8	3.0	15.7	20.6	11.2	2.2	7.5	4.6	4.0	0.27
	CoOp (Zhou et al. 2022)	34.6	9.3	12.3	2.8	52.1	31.5	28.9	13.2	21.0	4.6	5.5	0.70
	CSP (Nayak, Yu, and Bach 2023)	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.20
	HPL (Wang et al. 2023a)	46.4	18.9	19.8	6.9	63.4	48.1	40.2	24.6	30.1	5.8	7.5	1.37
	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	2.40
	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	2.70
	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	2.68
	CAILA (Zheng, Zhu, and Nevatia 2024)	**51.0**	20.2	21.6	8.2	67.8	59.7	49.4	32.8	43.9	8.0	11.5	3.08
	I2CD (Ours)	50.6	20.6	21.4	8.2	**69.1**	**62.1**	**50.2**	**35.6**	**44.0**	**10.9**	**14.9**	**4.25**

Table 2: Performance comparison under the open-world setting.

compositions increases from 2k to nearly 300k in the open-world scenario, making the task substantially more challenging.

For performance evaluation, we follow the generalized CZSL protocol (Purushwalkam et al. 2019), where both seen and unseen compositions are included in the test set. The evaluation metrics include best seen accuracy (S), best unseen accuracy (U), harmonic mean (HM) of seen and unseen accuracies, and the area under the curve (AUC) of the seen-unseen accuracy curve. The AUC is particularly crucial as it provides a comprehensive measure of the model's ability to balance performance across seen and unseen pairs. We vary the bias from $-\infty$ to $+\infty$ to obtain the seen-unseen accuracy curve, selecting models based on the best AUC observed on the validation set and reporting their performance on the test set. Additionally, the best seen and unseen accuracies are reported by adjusting the bias to $-\infty$ and $+\infty$, respectively.

Implementation Details

Our model uses the pre-trained CLIP ViT-L/14 variant (Radford et al. 2021), with lightweight adapters inserted into the transformer encoder layers of CLIP to enable efficient fine-tuning while keeping the pre-trained parameters frozen (Zheng, Zhu, and Nevatia 2024). Following previous work (Esser, Rombach, and Ommer 2020), the invertible network T is constructed by stacking 12 layers of blocks consisting of ActNorm, Affine Coupling, and Shuffling layers. The transformation function in the Affine Coupling is a 3-layer MLP with a hidden dimension of 512. The correlation coefficient in the disentangled invertible network is set to 0.975, with the dimensions of the state and object components set to 384. The state projection and object projection networks are implemented as 3-layer MLPs. The training of the model combines composition prediction, disentanglement, and intervention losses, using the Adam optimizer (Kingma and Ba 2015) with a batch size of 32. The learning rates are set to 1×10^{-5} for MIT-States and C-GQA, and 2×10^{-5} for UT-Zappos. Data augmentation follows the strategy used in CSP (Nayak, Yu, and Bach 2023).

Comparisons

The comparison methods can be classified into two main categories: methods that utilize CLIP as the visual and language

Setting		MIT-States				UT-Zappos				C-GQA			
		S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
Closed-World	Full	**50.6**	**54.1**	**40.0**	**23.3**	**69.1**	**77.1**	**57.9**	**46.4**	44.1	**41.2**	**35.0**	**16.3**
	w/o Intervention	50.3	53.8	39.2	22.6	66.2	75.8	56.5	44.0	**45.2**	40.9	32.3	15.7
	w/o Disentanglement	49.6	52.6	39.0	22.1	69.0	73.7	55.7	43.2	44.5	40.0	32.0	15.0
Open-World	Full	**50.6**	**20.6**	**21.4**	**8.2**	**69.1**	**62.1**	**50.2**	**35.6**	44.0	**11.0**	**14.4**	**4.18**
	w/o Intervention	50.3	19.5	20.6	7.7	69.0	58.0	49.6	34.6	**45.2**	10.3	13.4	3.94
	w/o Disentanglement	49.0	19.1	19.8	7.1	68.4	60.1	47.6	33.3	44.3	9.3	12.8	3.58

Table 3: The results of ablation studies in both closed-world and open-world settings

encoder (w CLIP), and those that do not (w/o CLIP). The CLIP-based approaches include CoOp (Zhou et al. 2022), CSP (Nayak, Yu, and Bach 2023), HPL (Wang et al. 2023a), DFSP (Lu et al. 2023), Troika (Huang et al. 2024), CDS-CZSL (Li et al. 2024), and CAILA (Zheng, Zhu, and Nevatia 2024), etc. The non-CLIP-based methods encompass CGE (Naeem et al. 2021), Co-CGE (Mancini et al. 2022), KG-SP (Karthik, Mancini, and Akata 2022), DRANet (Li et al. 2023), and ADE (Hao, Han, and Wong 2023), etc.

Comparisons under Closed-World Setting. Results under the closed-world setting are illustrated in Table 1. The proposed I2CD model demonstrates statistically significant improvements across multiple evaluation metrics for all examined datasets. Notably, the model achieves higher accuracy rates on unseen compositions across three datasets, highlighting its robust generalization capabilities.

Comparisons under Open-World Setting. The open-world setting presents a more challenging scenario, requiring models to generalize across a vast array of unseen compositions. Table 2 presents a comparative analysis of the proposed method against several baseline approaches. In the C-GQA dataset, the proposed method achieves 13.4% in Harmonic Mean (HM) and 3.76% in Area Under Curve (AUC), establishing new state-of-the-art results across all evaluation metrics. These improvements are particularly significant given that C-GQA represents the most challenging dataset among the three examined, thus validating the method's effectiveness.

Ablation Study

To evaluate the critical components of the I2CD model, ablation studies were conducted examining both the intervention and disentanglement mechanisms. Table 3 presents the comparative results across three datasets: MIT-States, UT-Zappos, and C-GQA. The ablation study reveals several key findings. The significant performance contribution of both intervention and disentanglement components validates the effectiveness of the disentangle-compose-disentangle mechanism. While the removal of the intervention component results in modest performance degradation, the model maintains relatively high seen and unseen accuracies on C-GQA, despite decreases in Harmonic Mean (HM) to 32.3% and Area Under Curve (AUC) to 15.7% in the closed-world setting. These results suggest that although intervention enhances overall performance, the model exhibits inherent robustness even without this component.

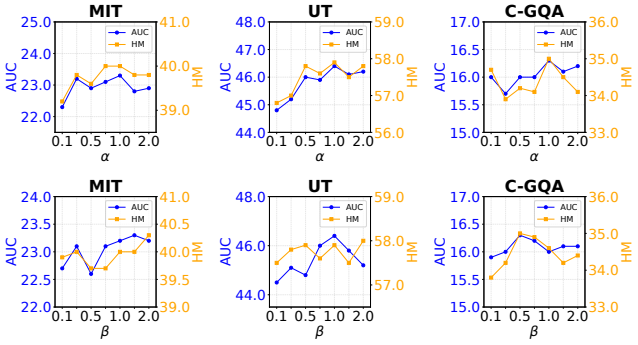

Figure 3: Hyper-parameter sensitivity analysis for α and β on the MIT-States, UT-Zappos, and C-GQA datasets.

Hyper-parameter Sensitivity Analysis

We conducted a sensitivity analysis on the hyper-parameters α and β by varying one parameter at a time while keeping all others fixed, with the results shown in Figure 3. For α, which controls the weight of the disentanglement loss, the best performance was consistently observed at $\alpha = 1.0$ across all datasets. For β, which regulates the causal intervention loss, the optimal values varied by dataset: $\beta = 2.0$ for MIT-States, $\beta = 1.0$ for UT-Zappos, and $\beta = 0.5$ for C-GQA. While these results highlight the need for dataset-specific tuning of β, the performance remained robust with minimal fluctuations across different values. Overall, the analysis confirms that the proposed method achieves robust performance under a wide range of hyper-parameter settings.

Conclusion

In this paper, we propose an invertible causal framework via disentangle-compose-disentangle (I2CD) for CZSL. The framework employs a three-stage mechanism for counterfactual generation, ensuring that modifications to individual primitives maintain mutual independence, thus facilitating robust causal disentanglement. Comprehensive experimental evaluations demonstrate the superior performance of I2CD across multiple benchmarks. Future research directions include investigating the contextual relationships between objects and states, an aspect beyond the scope of the current study.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos: 61802053, 62372387, 62402402, 62001400, 52441801), Natural Science Foundation of Sichuan Province (Grant No. 2024NSFSC0508), Sichuan Science and Technology Program (Grant No. 2024NSFSC0494), Fundamental Research Funds for the Central Universities (2682024ZTPY044, 2682025ZD004), China Postdoctoral Science Foundation (Grant No. 2021M702713), Special Research Funding under Yibin Municipal-University Dual Agreement (No. YBSCXY2024010012, YBSCXY2024010006), and the Fund of National Laboratory on Adaptive Optics, China, (Grant No. FNLAO-24-ZD-O02).

References

Atzmon, Y.; Kreuk, F.; Shalit, U.; and Chechik, G. 2020. A causal view of compositional zero-shot recognition. In *Adv. Neural Inform. Process. Syst.*

Bao, W.; Chen, L.; Huang, H.; and Kong, Y. 2024. Prompting Language-Informed Distribution for Compositional Zero-Shot Learning. In *Eur. Conf. Comput. Vis.*, 107–123.

Bengio, Y.; Deleu, T.; Rahaman, N.; Ke, N. R.; Lachapelle, S.; Bilaniuk, O.; Goyal, A.; and Pal, C. J. 2020. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. In *Int. Conf. Learn. Represent.*

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Int. Conf. Learn. Represent.*

Esser, P.; Rombach, R.; and Ommer, B. 2020. A disentangling invertible interpretation network for explaining latent representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 9223–9232.

Hao, S.; Han, K.; and Wong, K.-Y. K. 2023. Learning Attention as Disentangler for Compositional Zero-shot Learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Hu, X.; and Wang, Z. 2023. Leveraging sub-class discimination for compositional zero-shot learning. In *AAAI Conf. Artif. Intell.*, volume 37, 890–898.

Huang, S.; Gong, B.; Feng, Y.; Zhang, M.; Lv, Y.; and Wang, D. 2024. Troika: Multi-path cross-modal traction for compositional zero-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 24005–24014.

Isola, P.; Lim, J. J.; and Adelson, E. H. 2015. Discovering states and transformations in image collections. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1383–1391.

Jiang, C.; Wang, S.; Long, Y.; Li, Z.; Zhang, H.; and Shao, L. 2024a. Imaginary-Connected Embedding in Complex Space for Unseen Attribute-Object Discrimination. *IEEE Trans. Pattern Anal. Mach. Intell.*

Jiang, C.; and Zhang, H. 2024. Revealing the Proximate Long-Tail Distribution in Compositional Zero-Shot Learning. In *AAAI Conf. Artif. Intell.*, 2498–2506.

Jiang, D.; Chen, H.; Jing, H.; Ma, Y.; and Zheng, N. 2024b. MRSP: Learn Multi-representations of Single Primitive for Compositional Zero-Shot Learning. In *Eur. Conf. Comput. Vis.*

Karthik, S.; Mancini, M.; and Akata, Z. 2022. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 9336–9345.

Khan, M. G. Z. A.; Naeem, M. F.; Gool, L. V.; Pagani, A.; Stricker, D.; and Afzal, M. Z. 2023. Learning Attention Propagation for Compositional Zero-Shot Learning. In *IEEE Winter Conf. Appl. Comput. Vis.*, 3817–3826.

Kim, H.; Lee, J.; Park, S.; and Sohn, K. 2023. Hierarchical visual primitive experts for compositional zero-shot learning. In *Int. Conf. Comput. Vis.*, 5675–5685.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Int. Conf. Learn. Represent.*

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behav. Brain Sci.*, 40.

Li, X.; Yang, X.; Wei, K.; Deng, C.; and Yang, M. 2022a. Siamese Contrastive Embedding Network for Compositional Zero-Shot Learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 9326–9335.

Li, Y.; Liu, Z.; Chen, H.; and Yao, L. 2024. Context-based and Diversity-driven Specificity in Compositional Zero-Shot Learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 17037–17046.

Li, Y.; Liu, Z.; Jha, S.; and Yao, L. 2023. Distilled Reverse Attention Network for Open-world Compositional Zero-Shot Learning. In *Int. Conf. Comput. Vis.*, 1782–1791.

Li, Y.; Xu, Y.; Xu, X.; Mao, X.; and Lu, C. 2022b. Learning Single/Multi-Attribute of Object With Symmetry and Group. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12): 9043–9055.

Lu, X.; Guo, S.; Liu, Z.; and Guo, J. 2023. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 23560–23569.

Mancini, M.; Naeem, M. F.; Xian, Y.; and Akata, Z. 2021. Open world compositional zero-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 5222–5230.

Mancini, M.; Naeem, M. F.; Xian, Y.; and Akata, Z. 2022. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(3): 1545–1560.

Misra, I.; Gupta, A.; and Hebert, M. 2017. From red wine to red tomato: Composition with context. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1792–1801.

Naeem, M. F.; Xian, Y.; Tombari, F.; and Akata, Z. 2021. Learning graph embeddings for compositional zero-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 953–962.

Nagarajan, T.; and Grauman, K. 2018. Attributes as operators: factorizing unseen attribute-object compositions. In *Eur. Conf. Comput. Vis.*, 169–185.

Nayak, N. V.; Yu, P.; and Bach, S. H. 2023. Learning to Compose Soft Prompts for Compositional Zero-Shot Learning. In *Int. Conf. Learn. Represent.*

Purushwalkam, S.; Nickel, M.; Gupta, A.; and Ranzato, M. 2019. Task-driven modular networks for zero-shot compositional learning. In *Int. Conf. Comput. Vis.*, 3593–3602.

Qu, H.; Wei, J.; Shu, X.; and Wang, W. 2025. Learning Clustering-based Prototypes for Compositional Zero-shot Learning. *CoRR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 8748–8763.

Reddy, A. G.; L, B. G.; and Balasubramanian, V. N. 2022. On Causally Disentangled Representations. In *AAAI Conf. Artif. Intell.*, 8089–8097.

Ruis, F.; Burghouts, G.; and Bucur, D. 2021. Independent prototype propagation for zero-shot compositionality. *Adv. Neural Inform. Process. Syst.*, 34: 10641–10653.

Saini, N.; Pham, K.; and Shrivastava, A. 2022. Disentangling Visual Embeddings for Attributes and Objects. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 13658–13667.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Towards Causal Representation Learning. *CoRR*, abs/2102.11107.

Suter, R.; Miladinovic, D.; Schölkopf, B.; and Bauer, S. 2019. Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness. In *Int. Conf. Mach. Learn.*, 6056–6065.

Wang, H.; Yang, M.; Wei, K.; and Deng, C. 2023a. Hierarchical Prompt Learning for Compositional Zero-Shot Recognition. In *Int. Joint Conf. Artif. Intell.*, 1470–1478.

Wang, Q.; Liu, L.; Jing, C.; Chen, H.; Liang, G.; Wang, P.; and Shen, C. 2023b. Learning Conditional Attributes for Compositional Zero-Shot Learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 11197–11206.

Xu, Z.; Cheng, D.; Li, J.; Liu, J.; Liu, L.; and Wang, K. 2023. Disentangled Representation for Causal Mediation Analysis. In *AAAI Conf. Artif. Intell.*, 10666–10674.

Yang, M.; Xu, C.; Wu, A.; and Deng, C. 2023. A Decomposable Causal View of Compositional Zero-Shot Learning. *IEEE Trans. Multimedia*, 25: 5892–5902.

Yu, A.; and Grauman, K. 2014. Fine-Grained Visual Comparisons with Local Learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 192–199.

Zhang, T.; Liang, K.; Zhang, K.; and Ma, Z. 2024. Learning Conditional Prompt for Compositional Zero-Shot Learning. In *Int. Conf. Multimedia and Expo*.

Zheng, Z.; Zhu, H.; and Nevatia, R. 2024. CAILA: Concept-Aware Intra-Layer Adapters for Compositional Zero-Shot Learning. In *IEEE Winter Conf. Appl. Comput. Vis.*, 1721–1731.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9): 2337–2348.