GOTZSL: Optimal Transport-Guided Graph-Aware Feature Alignment for Compositional Zero-Shot Learning

Anonymous ACL submission

Abstract

Compositional Zero-Shot Learning (CZSL) aims to recognize unseen attribute-object compositions by generalizing from seen ones. Existing prompt-based methods often suffer from textual feature shift, while graph-based approaches are limited by static structures and lack compositional adaptability.

We propose **GOTZSL**: Optimal Transport-Guided Graph-Aware Feature Alignment for Compositional Zero-Shot Learning, a unified framework that integrates triple prompt tuning, a graph-based adapter, and compositional visual adaptation. GOTZSL encodes state, object, and pair prompts through triple-level text templates, refines them via a compositional graph aligned with LLM-derived anchors, and disentangles LoRA-adapted visual features using a dual-branch MLP module.

To improve consistency and generalization, we introduce a pairwise *optimal transport loss* and *partial label smoothing* over semantically related classes. Evaluated on UT-Zappos, MIT-States, and CGQA under both closed- and open-world CZSL settings, GOTZSL achieves state-of-the-art performance, demonstrating robust compositional reasoning.

1 Introduction

Introduction. Inspired by human cognitive abilities, CZSL has become a central challenge in machine learning. CZSL aims to recognize novel compositions of familiar concepts, such as identifying *"ripe apple"* without having seen the specific combination during training, as illustrated in Figure 1. It requires two key abilities: (1) **generalization**, i.e., transferring knowledge to unseen attribute-object pairs; and (2) **relational reasoning**, i.e., understanding the semantic compatibility between attributes and objects (e.g., *"ripe"* is more likely to modify *"apple"* than *"rock"*).

Evaluation in CZSL is typically conducted under two settings: **Closed-World**, where test compositions are limited to a known set of unseen pairs, and **Open-World**, where predictions must be made over both seen and unseen compositions jointly, posing a more realistic and challenging scenario. These settings assess a model's ability to generalize and remain discriminative across a combinatorially large output space.

While recent advances have made progress in aligning vision and language representations, most methods treat compositions independently, ignoring the relational structure among concepts. In this work, we propose **GOTZSL**, a graph-aware framework that integrates structured knowledge into feature alignment, enabling robust and interpretable CZSL.

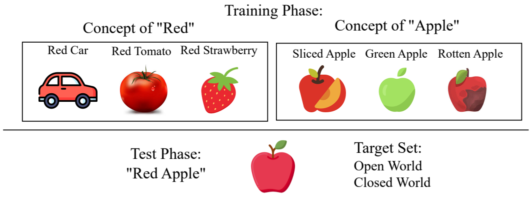

Figure 1: Illustration of the Compositional Zero-Shot Learning (CZSL) setting. The model is trained on individual attribute-object compositions (e.g., "Sliced Apple", "Rotten Apple"), but must recognize unseen combinations like "Red Apple" during testing.

Prior approaches to Compositional Zero-Shot Learning (CZSL) can be broadly categorized into three paradigms: (1) **Dual-classifier methods** (Li et al., 2020; Misra et al., 2017), which employ separate classifiers for states and objects but fail to capture their compositional dependencies; (2) **Graph-based methods**, such as (Mancini et al., 2021), which model state-object interactions via Graph Neural Networks (GNNs), or (Karthik et al., 2022), which leverage external knowledge to prune implausible compositions; and (3) **Semantic alignment methods** (Nagarajan and Grauman, 2018; Nan et al., 2019), which align visual and compositional embeddings by minimizing distance in a shared semantic space.

Despite recent advances, existing methods suffer from two key limitations: (i) they overlook the structural compatibility between attributes and objects, and (ii) they optimize visual and textual modalities in isolation, hindering generalization to unseen compositions. To address this, we propose a *graph-aware prompt tuning* strategy that models compositional hierarchies while preserving the semantic integrity of primitive concepts.

Inspired by these insights, we propose **GOTZSL**, a CLIP-based framework that combines soft prompting, graph-based adaptation, and multi-level alignment for CZSL. GOTZSL introduces triple prompt templates (state, object, composition) for the frozen CLIP text encoder and builds a compositional graph refined by a GCN using LLM-generated descriptions. To model visual primitives, LoRA adapters are injected into CLIP's image encoder for disentangling attribute and object features. Cross-modal alignment is achieved via contrastive losses across all axes, with predictions fused through a weighted strategy.

The contributions of this paper can be summarized as follows:

- First, we propose a novel graph adapter strategy that constructs a compositional graph to explicitly model relationships among state, object, and pair features. To enrich textual semantics, we incorporate LLM-generated sentences with diverse attribute-object combinations. Crucially, the attribute and object tokens appear in varying syntactic contexts.

- Second, we enhance CLIP's dual encoders by integrating multiple contextual soft prompt tokens and Low-Rank Adaptation (LoRA) modules. To improve alignment, we apply data augmentations to generate diverse visual primitives.

- Third, within the joint feature space, we employ optimal transport–guided objectives to align multi-branch predictions across attributes, objects, and pairs. To further enhance compositional consistency and mitigate overconfidence, we introduce partial label smoothing over semantically related classes.

The implementation will be made publicly available upon acceptance.

2 Related Work

Compositional Zero-Shot Learning (CZSL) without Pretrained VLMs. Unlike traditional ZSL (Guo and Guo, 2020; Li et al., 2021), which maps global class-level attributes, CZSL requires disentangling and recombining semantic primitives. Earlier CZSL methods fall into four categories: (1) *Dual-branch models* separately predict attributes and objects and combine results at inference (Li et al., 2022b; Yang et al., 2023a), but lack holistic modeling. (2) *Transform-based methods* learn transitions between compositions (**?**), yet rely heavily on transformation design. (3) *Joint embedding models* map visual features and composed concepts into a shared space (Purushwalkam et al., 2019), often at the cost of disentanglement. (4) *Graph-based approaches* (Naeem et al., 2021; Ge et al., 2022; Li et al., 2022a; Guo and Guo, 2023) use GNNs to encode relations among primitives and compositions, but usually rely on static graphs and fixed textual features.

Prompt Learning for CZSL. Prompt tuning has been recently explored to adapt pretrained vision-language models (VLMs), such as CLIP (Radford et al., 2021a), for CZSL. Methods like CSP (Nayak et al., 2022b) and DFSP (Lu et al., 2023) inject compositional prompts (e.g., "a photo of a sliced apple") to align visual and textual spaces. However, early approaches suffer from joint training collapse and primitive imbalance, limiting generalization.

Recent methods introduce structure-aware prompting. Hierarchical (Huynh and Elhamifar, 2023) and conditional prompts (Kang et al., 2023) improve disentanglement and adaptability. Other works (Yang et al., 2023b; Zhang et al., 2023; Jeong et al., 2023) leverage visual or linguistic context to refine prompts. While effective, these methods still struggle to fully disentangle semantic primitives and model their interactions robustly.

Graph-based Prompt Integration. Recent advances attempt to integrate graph reasoning with VLMs. Works like (Guo and Guo, 2023; Ge et al., 2022) use GCNs over CLIP embeddings to inject relational priors, but often rely on static graphs or hand-crafted text features. These approaches typically overlook the potential of learnable prompts and compositional semantics, motivating our approach to unify graph-based reasoning and prompt learning for CZSL.

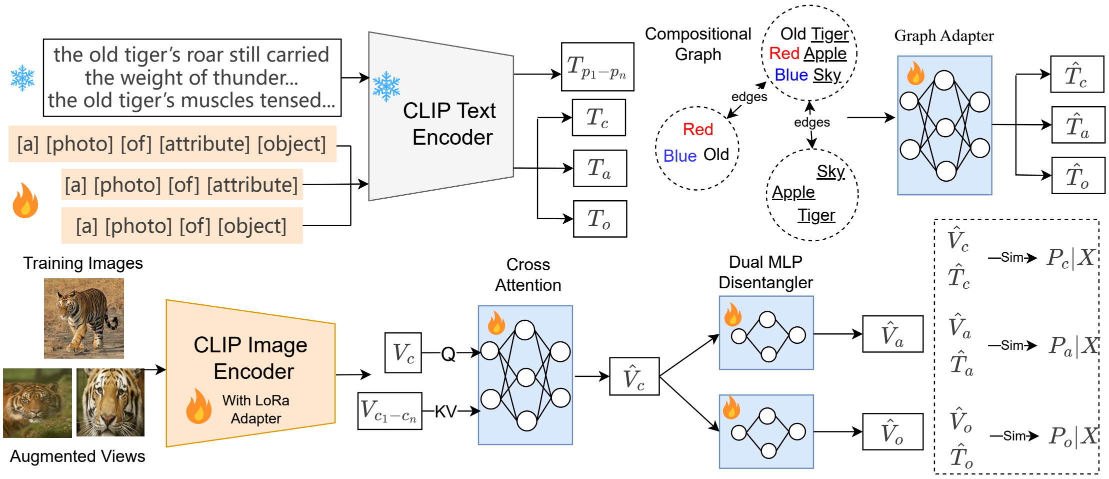

Figure 2: The overall architecture of GOTZSL. The model integrates three core components: (1) **Textual Prompt Encoding**—learnable soft prompts for states, objects, and their compositions are injected into CLIP's text encoder, forming a semantic graph refined by a Graph Convolutional Network (GCN) to capture contextual relationships; (2) **Visual Feature Adaptation**—the CLIP image encoder is adapted via LoRA and decomposed into state- and object-specific features through dual-branch extractors; (3) **Cross-Modal Alignment and Supervision**—visual and textual embeddings are aligned using contrastive loss, while an OT-based consistency loss enforces alignment between prediction logits from multiple branches.

3 Methodology

We introduce GOTZSL, a novel framework that advances Compositional Zero-Shot Learning (CZSL) by integrating: *(i) triple soft prompts for the text encoder*, *(ii) feature adaptation and decomposition for the image encoder*, and *(iii) graph-structured learning over state-object-pair compositions*. GOTZSL addresses the core challenge of generalizing to unseen attribute-object pairs through the following components:

- **Triple Soft Prompts with Graph Adaptation.** We design learnable prompts for *state*, *object*, and *composed pair* concepts, which are injected into CLIP's text encoder. These prompts are structured into a semantic graph, where nodes represent compositional primitives and edges encode their relationships. A unified GCN propagates contextual information and optimizes node features through contrastive supervision, enhancing the discriminability of valid versus invalid compositions.

- **Visual Adaptation and Decomposition.** We apply Low-Rank Adaptation (LoRA) to the upper layers of CLIP's image encoder to enable lightweight fine-tuning. The adapted visual features are further disentangled via dual branches into attribute-specific and object-specific embeddings, facilitating fine-grained alignment with textual semantics in the shared space.

- **Hierarchical Cross-Modal Alignment.** We optimize a dual-objective: (i) a contrastive loss aligns visual and textual embeddings of state-object pairs in the latent space, and (ii) a novel optimal transport (OT)–based consistency loss aligns prediction logits across state, object, and pair branches. This two-level supervision enforces semantic coherence and improves generalization to novel compositions.

3.1 Text Encoder

Triple Soft Prompts for Hierarchical Encoding GOTZSL employs *compositional soft prompts* to jointly learn composed and decomposed text representations for Compositional Zero-Shot Learning. Following CSP (Nayak et al., 2022a), we tokenize and encode the attribute and object names from the dataset, with each concept mapped to a dedicated embedding vector. To construct compositional prompts, we design and fill 3 prompt templates: (1) “`a photo of [attribute] [object]`”, (2) “`a photo of [attribute]`, (3) *“a photo of [object]”*. These templates generate triplet-level representations that disentangle and capture attribute semantics, object categories, and their interactions. The resulting soft prompts are optimized during training to align with visual features through our feature alignment and adapter modules. Furthermore, we encode the template prefix "`a photo of`" into three separate learnable context vectors, each corresponding to the attribute, object, and pair branches, respectively. This design offers greater flexibility for triple alignment by allowing independent contextual adaptation for each semantic role. Following PLID (Bao et al., 2024), we leverage LLM (Zhang et al., 2022) to generate multiple natural language descriptions for each attribute-object pair. These sentences are then encoded to obtain fixed LLM-derived textual features, which serve as base representations for aligning the learnable compositional prompts.

Cross-Attention Alignment with Structured Textual Descriptions To improve the alignment between attribute-object pairs and their textual representations, we incorporate a cross-attention mechanism that fuses learnable prompts with structured base features derived from a LLM. Given a batch of attribute-object pairs (a_i, o_i), we extract their corresponding LLM-based embeddings $\mathbf{T}_b \in \mathbb{R}^{T \times B \times d}$, where T is the number of descriptive tokens per pair, B is the batch size, and d is the embedding dimension. In parallel, we construct a query tensor $\mathbf{T}_q \in \mathbb{R}^{1 \times B \times d}$ from learnable prompts or class-level tokens to attend over $\mathbf{T}_b$. Both query and base features are first normalized via LayerNorm. We then apply multi-head cross-attention, enabling the query to dynamically attend to semantically relevant information in the base descriptions:

$$\tilde{\mathbf{T}}_q = \mathbf{T}_q + \text{MHA}(\text{LN}(\mathbf{T}_q), \text{LN}(\mathbf{T}_b), \text{LN}(\mathbf{T}_b)) \tag{1}$$

$$\mathbf{T}_q^{\text{out}} = \tilde{\mathbf{T}}_q + \text{MLP}(\text{LN}(\tilde{\mathbf{T}}_q)) \tag{2}$$

This design allows the class-level query to selectively aggregate contextual signals from structured textual descriptions, thereby enriching the output $\mathbf{T}_q^{\text{out}}$ with fine-grained and composition-aware semantics.

Compositional Graph Construction and Graph Adapter To enrich attribute and object text embeddings with structured semantic priors, we construct a compositional graph that captures relationships among attributes, objects, and their compositions. Each node in the graph represents a unique attribute, object, or attribute-object pair. Edges are formed based on shared semantics: between an attribute and a pair sharing the same attribute, or between an object and a pair sharing the same object. Additionally, pair nodes referencing the same composition are also interconnected. This graph is processed by a lightweight *Graph Adapter*—a Graph Convolutional Network (GCN)—to propagate contextual information and refine learnable text embeddings via message passing.

We formalize the structure as a semantic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes representing compositional units, and $\mathcal{E}$ denotes the set of bidirectional edges encoding semantic associations. The node set $\mathcal{V}$ consists of three types: (1) Attribute nodes ($\mathcal{S}$), representing primitive state concepts (e.g., "wet", "spotted"); (2) Object nodes ($\mathcal{O}$), representing entity categories (e.g., "apple", "dog"); (3) Reference pair nodes ($\mathcal{C}_{\text{ref}}$), representing fixed attribute-object compositions derived from LLMs, serving as semantic anchors.

Edges in $\mathcal{E}$ capture both semantic and structural relationships between nodes. Specifically, edges are constructed under the following rules: (1) Between an attribute node and a pair node if they share the same attribute; (2) Between an object node and a pair node if they share the same object; (3) Between reference pairs and their corresponding attribute or object nodes.

This compositional graph enables joint reasoning over both learnable prompt-based features and fixed LLM-derived textual knowledge. Through message passing, the Graph Adapter propagates semantic context across nodes, enriching the prompt embeddings with composition-aware information and promoting better generalization to novel attribute-object pairs.

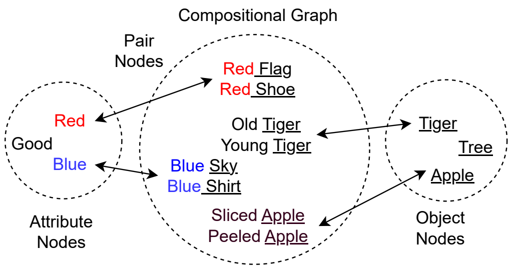

Figure 3: Visualization of the bidirectional compositional semantic graph.

Graph Adapter Formulation. Given node features for attributes $\mathbf{A} \in \mathbb{R}^{|\mathcal{A}| \times d}$, objects $\mathbf{O} \in \mathbb{R}^{|\mathcal{O}| \times d}$, and attribute-object pairs $\mathbf{P} \in \mathbb{R}^{|\mathcal{P}| \times d}$, we first normalize each type and concatenate to form the graph input:

$$\mathbf{X} = \begin{bmatrix} \hat{\mathbf{A}} \\ \hat{\mathbf{O}} \\ \hat{\mathbf{P}} \end{bmatrix} \in \mathbb{R}^{(|\mathcal{A}|+|\mathcal{O}|+|\mathcal{P}|)\times d}$$

We define an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where for each pair $(a_i, o_i) \in \mathcal{P}$, we add edges between the pair node and its attribute node a_i, as well as its object node o_i. The graph is processed by a 1-layer GCN (Kipf and Welling, 2016), denoted GCNConv, yielding updated node features:

$$\mathbf{X}' = \mathrm{GCNConv}(\mathbf{X}, \mathcal{E}) \tag{3}$$

We apply residual smoothing to attribute and object nodes for enhanced generalization:

$$\mathbf{A}' = \mathbf{A} + \mathbf{X}'_{\mathcal{A}}, \quad \mathbf{O}' = \mathbf{O} + \mathbf{X}'_{\mathcal{O}} \tag{4}$$

where $\mathbf{X}'_{\mathcal{A}}$ and $\mathbf{X}'_{\mathcal{O}}$ denote the corresponding slices of $\mathbf{X}'$. The refined features $\mathbf{A}', \mathbf{O}'$ are used in downstream compositional alignment.

3.2 Image Encoder

Integrating LoRA Adapter and Multi-View Encoding To enable efficient fine-tuning, we incorporate Low-Rank Adaptation (LoRA) modules into selected self-attention and feedforward layers of the image encoder. This lightweight design allows the model to adapt to novel attribute-object compositions with minimal memory and computational overhead. Given a batch of composed images $\mathbf{X} \in \mathbb{R}^{B\times V\times C\times H\times W}$, where V denotes the number of views per instance (typically 3), we designate the first as the anchor and generate two augmented views using weak transformations such as color jitter, contrast adjustment, and flipping.

$$\mathbf{V}_{\text{aug}} = \mathrm{ImgEncode}([\mathbf{X}_{\text{anchor}}, \mathbf{X}_1, \mathbf{X}_2]) \tag{5}$$

The resulting multi-view features $\mathbf{V}_{\text{aug}}$ facilitate robust learning of state-object interactions by capturing diverse appearance variations across views.

Aligning Composed Visual Features with Multi-View Contexts To enhance compositional reasoning, we align composed visual features with context embeddings extracted from multiple augmented views. View-specific representations are fused via cross-attention, enabling the composed embedding to integrate both local and global semantics for improved disambiguation of complex compositions. Given raw visual features from the anchor view $\mathbf{V}_{\text{anchor}}$ and two augmented views $\mathbf{V}^{(1)}_{\text{views}}, \mathbf{V}^{(2)}_{\text{views}}$, we perform cross-view enhancement as:

$$\mathbf{V}_{\text{pair}} = \mathbf{V}_{\text{anchor}} + \mathrm{CA}(\mathbf{V}_{\text{anchor}}, [\mathbf{V}^{(1)}_{\text{views}}, \mathbf{V}^{(2)}_{\text{views}}]) \tag{6}$$

This operation enriches the anchor view with complementary context, reinforcing the composed visual representation with cross-view semantics.

Extracting State and Object Representations from Composed Visual Features Given a composed visual representation, we disentangle it into separate attribute-sensitive and object-sensitive components. Two parallel MLP heads are applied to extract the corresponding features, yielding a triple representation: $\left[\mathbf{V}_{\text{pair}}, \mathbf{V}_{\text{attr}}, \mathbf{V}_{\text{obj}}\right]$. This decomposition enables independent alignment with attribute and object textual embeddings and improves interpretability in compositional recognition.

$$\mathbf{V}_{\text{attr}} = \mathrm{MLP}_{\text{attr}}(\mathbf{V}_{\text{pair}}), \mathbf{V}_{\text{obj}} = \mathrm{MLP}_{\text{obj}}(\mathbf{V}_{\text{pair}}) \tag{7}$$

3.3 Training

The overall training framework of GOTZSL is illustrated in Figure 2, which includes triple-branch text and image encoders, as well as modules for feature adaptation and extraction. In this section, we describe the computation of prediction logits and the corresponding training objectives.

Label Smoothing for CZSL. To address overconfidence and training instability arising from label sparsity and imbalanced primitive distributions, we apply *partial label smoothing* to the classification objectives for attributes, objects, and compositions. Unlike standard label smoothing, which uniformly redistributes confidence across all classes, our method selectively reallocates a portion of the probability mass to *semantically related* compositions. For each training instance, we construct a smoothing mask based on attribute or object overlap with the ground-truth label. For example, for the composition “sliced banana,” the smoothed target assigns high confidence to the correct class while also distributing some weight to compositions like “sliced apple” (same attribute) and “ripe banana” (same object). Concretely, the ground-truth receives confidence $(1 - \epsilon)$, and the remaining ϵ is distributed among related classes. This structured smoothing strategy enhances semantic consistency, mitigates overfitting, and improves generalization to unseen attribute-object combinations.

Compositional classification loss. A cross-entropy loss, optionally with partial label smoothing, is applied to the composed attribute-object pair logits to promote accurate compositional prediction. The logits are computed using CLIP similarity between visual and textual features: $\mathbf{z}_i^{\text{pair}} = \mathbf{v}_i^{\text{pair}} \cdot (\mathbf{t}^{\text{pair}})^\top$.

$$\mathcal{L}_{\text{pair}} = -\sum_{i=1}^{B} \mathbf{y}_i^{\text{pair}} \cdot \log\Big(\text{softmax}(\mathbf{z}_i^{\text{pair}})\Big) \tag{8}$$

Attribute classification loss. A cross-entropy loss with label smoothing supervising attribute classification. Attribute logits are computed as $\mathbf{z}_i^{\text{attr}} = \mathbf{v}_i^{\text{attr}} \cdot (\mathbf{t}^{\text{attr}})^\top$.

$$\mathcal{L}_{\text{attr}} = -\sum_{i=1}^{B} \mathbf{y}_i^{\text{attr}} \cdot \log\big(\text{softmax}(\mathbf{z}_i^{\text{attr}})\big) \tag{9}$$

Object classification loss. A cross-entropy loss with label smoothing applied to object classification. Object logits are calculated as $\mathbf{z}_i^{\text{obj}} = \mathbf{v}_i^{\text{obj}} \cdot (\mathbf{t}^{\text{obj}})^\top$.

$$\mathcal{L}_{\text{obj}} = -\sum_{i=1}^{B} \mathbf{y}_i^{\text{obj}} \cdot \log\Big(\text{softmax}(\mathbf{z}_i^{\text{obj}})\Big) \tag{10}$$

Pairwise Optimal Transport Loss To promote compositional generalization, we introduce a *pairwise optimal transport (OT) loss* with entropy regularization, which aligns the model's pairwise predictions with the joint distribution formed by its attribute and object predictions. Specifically, we interpret the outer product of attribute and object probability distributions as a soft joint prediction, and encourage consistency with the pairwise logits using the entropy-regularized Wasserstein distance computed via the Sinkhorn algorithm. Formally, the loss is defined as:

$$\mathcal{L}_{\text{OT}} = \sum_{i=1}^{B} \Big\langle T^{(i)}, \mathbf{D}^{(i)} \Big\rangle = \sum_{i=1}^{B} \sum_{j=1}^{|\mathcal{A}|} \sum_{k=1}^{|\mathcal{O}|} T_{jk}^{(i)} \cdot D_{jk}^{(i)} \tag{11}$$

where B is the batch size, $T^{(i)} \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{O}|}$ denotes the optimal transport plan for the i-th sample, and $\mathbf{D}^{(i)} = 1 - \mathbf{P}_{\text{attr}}^{(i)} \otimes \mathbf{P}_{\text{obj}}^{(i)}$ is the cost matrix based on the outer product between predicted attribute and object distributions. The transport plan $T^{(i)}$ is obtained using entropy-regularized Sinkhorn iterations for stable and efficient optimization.

This loss encourages the model to produce consistent compositional predictions across factorized and holistic outputs, enhancing its ability to generalize to unseen attribute-object combinations.

Total Training Loss. To learn disentangled and compositional representations for Compositional Zero-Shot Learning (CZSL), we optimize a multi-objective loss that supervises both individual components (attributes and objects) and their compositions. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pair}} + \mathcal{L}_{\text{attr}} + \mathcal{L}_{\text{obj}} + \lambda\, \mathcal{L}_{\text{OT}} \tag{12}$$

where λ is a hyperparameter that controls the strength of the optimal transport (OT) guidance.

Each component encourages the model to capture a distinct aspect of compositional structure: $\mathcal{L}_{\text{pair}}$ focuses on joint composition classification, $\mathcal{L}_{\text{attr}}$ and $\mathcal{L}_{\text{obj}}$ supervise disentangled attribute and object recognition, and $\mathcal{L}_{\text{OT}}$ provides fine-grained alignment signals between visual and textual modalities. This formulation ensures effective generalization to unseen attribute-object compositions during inference.

Inference with Multi-Branch Logits. During inference, the model predicts compositional concepts by combining semantic signals from three branches: *pair*, *attribute*, and *object*. Given the composed visual feature, we first obtain branch-specific logits: $\text{logits}_{\text{pair}}$, $\text{logits}_{\text{attr}}$, and $\text{logits}_{\text{obj}}$ via cosine similarity with their respective textual prototypes, following:

$$\text{logits} = \tau^{-1} \cdot \langle \hat{\mathbf{v}}, \hat{\mathbf{t}} \rangle = \tau^{-1} \cdot (\hat{\mathbf{v}}^\top \hat{\mathbf{t}}) \tag{13}$$

where $\hat{\mathbf{v}}$ and $\hat{\mathbf{t}}$ are ℓ_2-normalized visual and textual features. To improve compositional consistency, we integrate these predictions using a rule-based fusion strategy. Specifically, the final score for each candidate pair (a_i, o_i) is computed as a weighted combination of its direct pairwise logit and the product of independent attribute and object probabilities:

$$\text{logits}_{\text{final}}^{(a_i,o_i)} = \alpha \cdot \text{logits}_{\text{pair}}^{(a_i,o_i)} + \beta \cdot P_{\text{attr}}(a_i) \cdot P_{\text{obj}}(o_i) \tag{14}$$

where $P_{\text{attr}} = \text{softmax}(\text{logits}_{\text{attr}})$, $P_{\text{obj}} = \text{softmax}(\text{logits}_{\text{obj}})$, and (α, β) are weighting coefficients. This fusion encourages agreement between holistic and factorized predictions, enabling more robust inference over unseen compositions.

4 Experiments

Table 1: Dataset statistics and descriptions under the CZSL setting.

Dataset	#A	#O	Seen	Unseen	Description
UT-Zappos	16	12	83	33	Fine-grained shoes with subtle visual attributes.
MIT-States	115	245	1262	700	Natural images with diverse state-object pairs.
CGQA	117	150	16122	5536	GQA-based compositional set with rich attributes.

In this section, we comprehensively evaluate GOTZSL on three widely-used compositional vision benchmarks: UT-Zappos, MIT-States, and CGQA. These datasets span diverse visual domains, including fine-grained object classification (e.g., shoes in UT-Zappos), state-based recognition (e.g., verb-noun compositions in MIT-States), and texture-centric scenes (e.g., CGQA). We conduct experiments under both closed-world (CW) and open-world (OW) settings to assess the generalization ability of GOTZSL to unseen attribute-object compositions. A summary of dataset statistics and descriptions is provided in Table 1.

Training Details and Evaluation Metrics We fine-tune GOTZSL with ViT-L/14 (Radford et al., 2021a) as the CLIP backbone using AdamW (Kingma and Ba, 2014) and a step-based scheduler. Training runs for 20 epochs per dataset with a learning rate of 1×10^{-4}; batch size and other hyperparameters are dataset-dependent. All experiments are performed on a single NVIDIA A100 GPU.

We evaluate GOTZSL and recent CLIP-based CZSL baselines—e.g., TsCA (Li et al., 2024) and DCDA (Geng et al., 2025)—under both Closed-World (CW) and Open-World (OW) settings across three benchmarks: UT-Zappos, MIT-States, and CGQA. Metrics include Seen Accuracy (S), Unseen Accuracy (U), Harmonic Mean (H), and AUC. Results are shown in Table 2 and Table 3.

We report both quantitative metrics and qualitative insights on compositional disentanglement and generalization.

Closed-World Evaluation Summary GOTZSL consistently achieves strong performance across all benchmarks in the Closed-World setting. On UT-Zappos, it achieves the best results across all metrics, including a harmonic mean (H) of 60.0% and an AUC of 46.9%. On CGQA, GOTZSL sets a new state-of-the-art with a 34.8% H and 16.4% AUC, significantly outperforming prior methods. On MIT-States, it ranks second in seen accuracy (52.4%) and maintains competitive generalization with a 39.0% H.

These results highlight GOTZSL's effectiveness in compositional reasoning, driven by its ability to align visual and textual primitives across seen and unseen compositions. While DCDA shows strong performance on MIT-States and TsCA remains competitive across all datasets, GOTZSL exhibits overall superior balance and generalization.

Open-World Evaluation Summary In the more challenging Open-World setting, GOTZSL continues to perform competitively. On CGQA—the most difficult benchmark—it achieves the best unseen accuracy (11.4%), harmonic mean (15.6%), and AUC (4.6%), setting a new state-of-the-art. On UT-Zappos, GOTZSL maintains strong generalization with an H of 51.7% and AUC of 36.6%. On MIT-States, it delivers results comparable to the top-performing baselines.

These findings confirm that GOTZSL benefits from structured alignment and disentangled representations, enabling robust generalization to novel attribute-object compositions under both CW and OW settings.

4.1 Ablation Studies

To further validate the effectiveness of each module in GOTZSL, we conduct ablation studies on the UT-Zappos dataset, as reported in Table 4. The results show that removing any individual component leads to a degradation in overall performance, highlighting the importance of each module. Notably, removing the Graph Adapter results in the most significant drop in AUC (2.5%), underscoring the effectiveness of our proposed compositional graph structure in enhancing compositional generalization.

5 Conclusion

We present GOTZSL, a unified framework for compositional zero-shot learning that integrates rich semantic text encoding, multi-view visual encoding, disentangled representation learning, and structured semantic alignment. By leveraging graph-based reasoning and multi-branch prediction, GOTZSL achieves strong generalization to unseen

Table 2: Closed-World CZSL results on **UT-Zappos**, **MIT-States** and **C-GQA** datasets. Evaluation metrics include Seen Accuracy (S), Unseen Accuracy (U), Harmonic Mean (H), and Area Under the Curve (AUC). The best results are highlighted in **bold**, and the second-best results are underlined.

Method	MIT-States				UT-Zappos				C-GQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
CLIP (Radford et al., 2021b)	30.2	40.0	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 et al., 2022a)	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33	28.8	26.8	20.5	6.2
DSFP (Lu et al., 2023)	46.9	52	37.3	20.6	66.7	71.7	47.2	36	38.2	32	27.1	10.5
GIPCOL (Xu et al., 2024)	48.5	49.6	36.6	19.9	65	68.5	48.8	36.2	32	28.4	22.5	7.14
Troica (Nayak et al., 2022a)	49	53	39.3	22.1	66.8	73.8	54.6	41.7	41	35.7	29.4	12.4
CDS-CZSL (Nayak et al., 2022a)	50.3	52.9	39.2	22.4	63.9	74.8	52.7	39.5	38.3	34.2	28.1	11.1
PLID (Bao et al., 2024)	49.7	52.4	39.0	22.1	67.3	68.8	52.4	38.7	38.8	33	27.9	11
MSCI (Wang et al., 2025)	50.2	53.4	39.9	22.8	67.4	75.5	59.2	45.8	42.4	38.2	31.7	14.2
DCDA (Geng et al., 2025)	**57.1**	**55.5**	**43.1**	**27.0**	69.1	74.1	57.2	44.2	38.5	28.8	25.3	9.4
TsCA (Li et al., 2024)	51.2	52.9	39.9	23	68.7	**75.8**	58.5	46.1	43.8	38.9	33.1	15.2
GOTZSL(Ours)	52.4	52.0	39.0	22.9	**70.4**	75.0	**60.0**	**46.9**	**45.5**	**40.7**	**34.8**	**16.4**

Table 3: Open-World CZSL results on **UT-Zappos**, **MIT-States**, and **C-GQA** datasets. Evaluation metrics include Seen Accuracy (S), Unseen Accuracy (U), Harmonic Mean (H), and Area Under the Curve (AUC). The best results are highlighted in **bold**, and the second-best results are underlined.

Method	MIT-States				UT-Zappos				C-GQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
CLIP (Radford et al., 2021b)	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 et al., 2022a)	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.20
DSFP (Lu et al., 2023)	47.5	18.5	19.3	6.8	66.8	60	44	30.3	38.3	7.2	10.4	2.4
GIPCOL (Xu et al., 2024)	48.5	16	17.9	6.3	65	45	40.1	23.5	31.6	5.5	7.3	1.3
Troica (Nayak et al., 2022a)	48.8	18.7	20.1	7.2	66.4	61.2	47.8	33	40.8	7.9	10.9	2.7
CDS-CZSL (Nayak et al., 2022a)	49.4	21.8	22.1	8.5	64.7	61.3	48.2	32.3	37.6	8.2	11.6	2.7
PLID (Bao et al., 2024)	49.1	18.7	20	7.3	67.6	55.5	46.6	30.8	39.1	7.5	10.6	2.5
MSCI (Wang et al., 2025)	49.2	20.6	21.2	7.9	67.4	63.0	**53.2**	**37.3**	42	10.6	13.7	3.8
DCDA (Geng et al., 2025)	**55.0**	**27.7**	**26.7**	**12.0**	67.8	62.5	51.4	35.8	35.3	6.4	8.5	1.76
TsCA (Li et al., 2024)	50.8	21.7	22.3	8.7	**69.8**	**63.4**	52.2	37.1	**44.3**	**11.4**	14.7	4.5
GOTZSL(Ours)	49.8	21.5	21.9	8.5	67.3	62.5	51.7	36.6	44.0	**11.4**	**15.6**	**4.6**

Table 4: Ablation results for GOTZSL on UT-Zappos in the Close-World setting.

Ablation Experiment	S	U	H	AUC
w/o GraphAdapter	68.1	73.0	54.7	44.4
w/o LoRaAdapter	69.5	73.7	59.1	45.8
w/o Text CrossAttention	70.0	73.8	59.0	45.8
w/o Visual Disentanglers	69.0	73.6	58.8	46.4
w/o Visual CrossAttention	**71.0**	73.9	59.2	46.5
Ours (Full)	70.4	**75.0**	**60.0**	**46.9**

attribute-object compositions across diverse benchmarks. Our results demonstrate the effectiveness of combining graph-aware structured textual priors with fine-grained visual cues in compositional settings.

Limitations

Two key limitations are identified in GOTZSL: (1) The current fusion mechanism relies on fixed weights, which may limit adaptability to diverse visual domains; (2) The model assumes clean attribute-object annotations during training, which may not generalize well to noisy or weakly labeled data. Addressing these limitations presents opportunities for future work in adaptive fusion and weakly supervised compositional learning.

References

Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. 2024. Prompting language-informed distribution for compositional zero-shot learning. In *European Conference on Computer Vision*, pages 107–123. Springer.

Chuanyi Ge, Bin-Bin Huang, Jie Chen, and Min-Ling

Wang. 2022. Compositional zero-shot learning via multi-branch graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 686–694.

Yuxia Geng, Runkai Zhu, Jiaoyan Chen, Jintai Chen, Xiang Chen, Zhuo Chen, Shuofei Qiao, Yuxiang Wang, Xiaoliang Xu, and Sheng-Jun Huang. 2025. Graph-guided cross-composition feature disentanglement for compositional zero-shot learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2678–2690.

Jingcai Guo and Song Guo. 2020. A neural process approach for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5653–5662.

Jingcai Guo and Song Guo. 2023. Graph-based open-world compositional zero-shot learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14.

Dang Huynh and Ehsan Elhamifar. 2023. Hierarchical prompt learning for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19053–19063.

Jaehoon Jeong, Jinwoo Lee, Juho Kim, and Nojun Kwak. 2023. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 36.

Mingu Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825.

Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. 2022. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9336–9345.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Miaoge Li, Jingcai Guo, Richard Yi Da Xu, Dongsheng Wang, Xiaofeng Cao, Zhijie Rao, and Song Guo. 2024. Tsca: On the semantic consistency alignment via conditional transport for compositional zero-shot learning. *arXiv preprint arXiv:2408.08703*.

Xiangyu Li, Jingcai Guo, and Song Guo. 2022a. Embedding visual-semantic relationships into graph matching for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18401.

Xiangyu Li, Jingcai Guo, and Song Guo. 2022b. Siamese contrastive embedding network for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9335.

Xiangyu Li, Xu Yang, Kun Wei, and Cheng Deng. 2021. Generalized zero-shot learning via multi-view embedding network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1314–1323.

Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. 2020. Symmetry and group in attribute-object compositions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11316–11325.

Yuning Lu, Tianzhu Liu, Xiang Zhang, and Yongdong Zhang. 2023. Disentangled prompt tuning for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22796–22805.

Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. 2021. Learning graph embeddings for open world compositional zero-shot learning. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 44, pages 5226–5242.

Ishan Misra, Abhinav Gupta, and Martial Hebert. 2017. From red wine to red tomato: Composition with context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1792–1801.

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

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

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

Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. 2022a. Compositional prompt tuning with motion cues for open-vocabulary video relation detection. In *International Conference on Learning Representations*.

Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. 2022b. Learning compositional representations for effective low-shot generalization. In *International Conference on Learning Representations*.

Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. 2019. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3593–3602.

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

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

Yue Wang, Shuai Xu, Xuelin Zhu, and Yicong Li. 2025. Msci: Addressing clip's inherent limitations for compositional zero-shot learning. *arXiv preprint arXiv:2505.10289*.

Guangyue Xu, Joyce Chai, and Parisa Kordjamshidi. 2024. Gipcol: Graph-injected soft prompting for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5774–5783.

Muli Yang, Yizhou Wang, Fuzhen Zhuang, Haofen Wang, and Xian-Sheng Hua. 2023a. Dual complementary learning for open-world compositional zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2913–2922.

Muli Yang, Yizhou Wang, Fuzhen Zhuang, Haofen Wang, and Xian-Sheng Hua. 2023b. Prompting language-informed distribution for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14777–14786.

Chi Zhang, Guozheng Song, Yue Zhang, Baoquan Li, Yong Liu, and Fuzhen Zhuang. 2023. Visual-adaptive prompt tuning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2902–2912.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. In *International Journal of Computer Vision*, volume 130, pages 2337–2348.