MLDA: Test-Time Multi-Level Adaptation with Dynamic Alignment for Compositional Zero-Shot Learning

Miaoge Li , Yu Liu , Jingcai Guo*
Department of COMP/LSGI, The Hong Kong Polytechnic University, Hong Kong SAR
jc-jingcai.guo@polyu.edu.hk

Abstract

Compositional Zero-Shot Learning (CZSL) aims to recognize novel *attribute–object* compositions by leveraging shared knowledge from the seen primitives. While recent works have begun to incorporate test-time adaptation to mitigate performance degradation caused by label-space distribution shifts, they suffer from primitive-level knowledge underutilization and cross-modal semantic asymmetry, hindering robust understanding of novel compositions. To address these issues, we introduce MLDA, a test-time adaptation framework for CZSL that jointly derives multi-level prototypes and performs dynamic vision–language alignment. Specifically, we construct and progressively update composition- and primitive-level prototype sets to accumulate structured semantic knowledge. By augmenting test inputs with diverse visual perspectives and enriched compositional descriptions, MLDA leverages optimal transport to achieve consistent matching of semantically salient information across modalities. Furthermore, composition activation scores are dynamically updated over the test stream to emphasize informative compositions and suppress irrelevant ones. Extensive experiments are conducted to verify the effectiveness of the proposed method. The code is available at https://github.com/keepgoingjkg/MLDA.

1 Introduction

Humans exhibit a remarkable ability to recognize and reason about novel concepts by composing previously acquired knowledge [Lake *et al.*, 2017]. For instance, we can effortlessly identify an unfamiliar concept like `wooden laptop` by combining the known attribute `wooden` with the object `laptop`, even without prior exposure to this combination. This capacity for compositional generalization motivates Compositional Zero-Shot Learning (CZSL) [Misra *et al.*, 2017; Naeem *et al.*, 2021], which recognizes unseen attribute–object pairs by transferring knowledge from the seen ones.

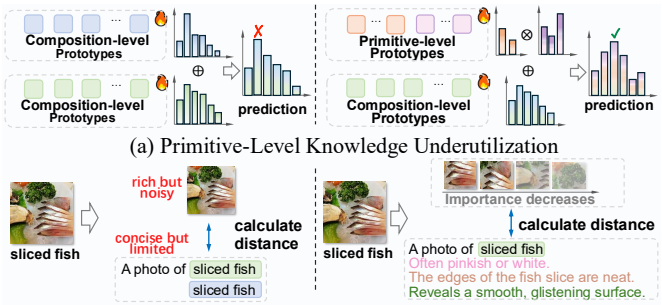


Figure 1: Motivation of MLDA. Existing CZSL methods (*left*) suffer from two key limitations at test time:(a) over-reliance on composition-level prototypes while neglecting primitive-level knowledge, and (b) point-to-point matching that fails to address cross-modal semantic asymmetry. Conversely, MLDA (*right*) adopts multi-level prototype adaptation and distribution-aware alignment, enabling complementary knowledge accumulation and robust cross-modal matching.

The advent of large-scale pre-trained vision–language models (VLMs) has led to a series of prompt-tuning methods that adapt these models to the CZSL setting in a training-based manner, leveraging the versatile multi-modal representations encoded in VLMs [Nayak *et al.*, 2023; Lu *et al.*, 2023; Huang *et al.*, 2024]. However, CZSL inherently suffers from distribution shifts caused by unseen combinations of sub-concepts at test time. This challenge has recently motivated test-time adaptation for CZSL, which aims to dynamically adjust model behavior when encountering novel compositions. Existing approaches include test-time prompt tuning for composition labels [Zhou and Ma, 2024] or multi-modal prototype learning with semantic consistency constraints [Yan and Feng, 2025], both of which achieve encouraging performance gains. Nevertheless, current test-time adaptation methods for CZSL remain limited in two key aspects: *(1) Primitive-Level Knowledge Underutilization.* These methods typically learn only one or a few composition-level prototypes and fail to accumulate knowledge about individual primitives as test samples are processed. *(2) Cross-Modal Semantic Asymmetry.* In practice, visual inputs often contain rich but noisy information. For example, as shown in Fig.1(b), an image of `sliced fish` may also include irrelevant items, such as broccoli. In contrast, composition prototypes encode concise but limited semantic

*Corresponding author: *Jingcai Guo*.

content. This inherent heterogeneity between visual and textual representations induces cross-modal semantic imbalance, which undermines generalization to novel compositions.

To address the above issues, we propose a test-time **M**ulti-**L**evel Adaptation with **D**ynamic **A**lignment (dubbed MLDA) framework for CZSL. Initially, moving beyond reliance on composition-level prototypes alone, we construct comprehensive prototype representations at both the composition and primitive levels. This multi-level design enables MLDA to not only model each composition more precisely, but also accumulate reliable primitive-level knowledge as test samples are continuously observed. By explicitly integrating complementary side information from different granularities, MLDA effectively mitigates error propagation caused by unreliable composition predictions and improves both accuracy and robustness.

Moreover, to facilitate fine-grained alignment between visual and textual domains, given a candidate composition and a test image, MLDA augments the visual input via image transformations and generates a set of potential compositional descriptions using large language models (LLMs) [Brown *et al.*, 2020; Ouyang *et al.*, 2022; Wang *et al.*, 2024]. Each enhanced view or description captures a specific semantic aspect of the input, and together they provide a more comprehensive understanding of the underlying composition. Intuitively, not all augmentations contribute equally to composition recognition, as some may correspond to irrelevant visual regions or imprecise textual descriptors. To this end, MLDA dynamically weights the importance of each augmentation according to its prediction entropy, allowing informative cues to be emphasized while suppressing noisy ones. More importantly, unlike existing methods that rely on point-to-point distance matching, MLDA represents the augmented visual instances and compositional descriptions as two discrete distributions and leverages optimal transport (OT) [Villani and others, 2008] to perform alignment at the distribution level. This formulation explicitly models inter-modal interactions and enables consistent matching of semantically relevant information.

In terms of data distribution, the composition space is inherently imbalanced and partially infeasible [Jiang and Zhang, 2024]. During test-time adaptation, only a subset of candidate compositions is relevant to the current test stream. To capture this property, MLDA assigns each composition to a dynamically updated activation score that reflects its test-time relevance. First, the activation score serves as a gating mechanism that determines whether a composition participates in OT–based alignment, enabling the model to focus on confident and valid candidates rather than blindly operating over the entire composition space. Second, it reweights the aggregated outputs, allowing more reliable candidates to exert greater influence on the final prediction.

In summary, our contributions are fourfold:

- We introduce multi-level adaptation for CZSL in the test-time scenario. By learning prototypes at both the composition and the primitive levels, our approach enables more accurate understanding of compositions.
- We leverage informative multi-modal knowledge to perform fine-grained semantic alignment of key concepts by formulating the image-text distance calculation as an optimal transport problem.
- We propose composition activation scores that gate OT-based alignment and reweight predictions, enabling the model to dynamically prioritize reliable candidates.
- Extensive comparisons and ablation studies on three benchmarks demonstrate the effectiveness of the proposed MLDA.

2 Related Work

2.1 Compositional Zero-Shot Learning

Compositional Zero-Shot Learning (CZSL) aims to recognize unseen compositions by generalizing from seen ones, operating within a fixed set of known attributes and objects [Mancini *et al.*, 2021]. Prior research in CZSL can be broadly categorized into two paradigms: compositional recognition and primitive recognition. The former learns compositional representations directly, either by aligning visual features with textual labels in a shared embedding space or by leveraging graph-based models to capture contextual relationships [Tanwisuth *et al.*, 2023; Saini *et al.*, 2022; Li *et al.*, 2022]. The latter independently predicts sub-concepts using dual classifiers, and emphasize disentangling visual representations through strategies such as contrastive learning [Li *et al.*, 2022] and knowledge distillation [Li *et al.*, 2023].

Advances in large-scale vision–language models (VLMs) like CLIP have increasingly enabled CZSL to leverage semantic knowledge from pre-trained multimodal representations [Wang *et al.*, 2023b; Li *et al.*, 2024b; Rao *et al.*, 2025]. Several works adopt prompt learning and formulate CZSL as a multi-path identification problem, enabling composition reasoning through diverse semantic perspectives. For instance, Troika [Huang *et al.*, 2024] employs a cross-modal traction module to align branch-specific prompt representations and adjusts them based on visual content. TsCA [Li *et al.*, 2024a] builds three semantically homologous sets and enforces fine-grained alignment with a cycle-consistency constraint. CLUSPRO [Qu *et al.*, 2025] learns multiple prototypes for each primitive via within-primitive clustering and dynamically updates them to capture diversity. Despite notable progress, existing CZSL methods generally keep model parameters fixed after training, leaving test-time adaptation largely unexplored.

2.2 Test-Time Adaptation

Test-time adaptation (TTA) is an effective technique to address distribution shifts between training and test data [Wang *et al.*, 2020; Nado *et al.*, 2020; Khurana *et al.*, 2021]. It is particularly beneficial for real-world deployment in areas such as autonomous driving and medical diagnosis [Karmanov *et al.*, 2024]. For example, TENT [Wang *et al.*, 2020] adapts the affine parameters of batch normalization layers by minimizing the entropy of prediction probabilities. MEMO [Zhang *et al.*, 2022] enforces invariant predictions across different augmentations of each test sample. Motivated by the success of VLMs, recent studies adopt test-time prompt tuning (TPT) to optimize textual prompts via entropy minimization. Specifically, TPT [Shu *et al.*, 2022] and DiffTPT [Feng *et al.*, 2023] adapt pre-trained models by learning domain-specific prompts

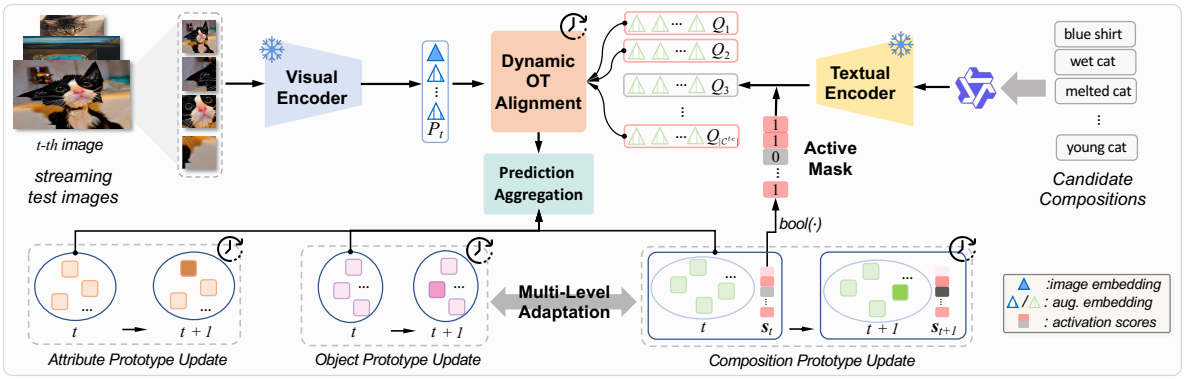


Figure 2: The overall framework of the proposed MLDA (zoom-in for more details).

from test data. DynaPrompt [Xiao *et al.*, 2025] introduces an online prompt buffer to dynamically select and optimize prompts relevant to incoming test samples.

More recently, TTA has been extended to the CZSL setting. Zhou et al. [2024] provide the first exploration by applying visual augmentations to test images and minimizing the entropy of the resulting prediction distributions. TOMCAT [Yan and Feng, 2025] further proposes to accumulate knowledge from both visual and textual modalities to update multimodal prototypes at test time. Despite their effectiveness, these approaches primarily operate at the composition level, fail to explicitly model fine-grained semantic matching, and are sensitive to noise in test data.

3 Methodology

In this section, we first briefly review the preliminaries of CZSL. We then introduce the details of our MLDA, which performs multi-level adaptation via composition- and primitive-level prototype updating, together with dynamic semantic alignment and prediction aggregation at test time. An overview of the proposed framework is shown in Fig. 2.

3.1 Preliminaries

Given the attribute set $\mathcal{A} = \{a_0, a_1, \ldots, a_{|\mathcal{A}|}\}$ and object set $\mathcal{O} = \{o_0, o_1, \ldots, o_{|\mathcal{O}|}\}$, the composition label space is defined as their Cartesian product, i.e., $\mathcal{C} = \mathcal{A} \times \mathcal{O}$. This space can be partitioned into two disjoint subsets: the seen subset $\mathcal{C}^{se}$ and the unseen subset $\mathcal{C}^{us}$, with $\mathcal{C}^{se} \cap \mathcal{C}^{us} = \emptyset$. During training, only $\mathcal{C}^{tr} = \mathcal{C}^{se}$ is used to learn a discriminative model $\mathcal{M}^{tr}$ that maps input images to compositional labels. During testing, the trained $\mathcal{M}^{tr}$ is expected to predict compositions of test samples X_{test} in the test space $\mathcal{C}^{te}$. Concretely, in the closed-world setting, $\mathcal{C}^{te} = \mathcal{C}^{se} \cup \mathcal{C}^{us}$, whereas in the open-world setting, all possible attribute–object pairs are considered, i.e., $\mathcal{C}^{te} = \mathcal{C}$. In contrast, under test-time adaptation for CZSL, an adaptive model $\mathcal{M}^{ada}$ has access only to unlabeled test samples and updates its predictions online to better handle the distribution shift between $\mathcal{C}^{tr}$ and $\mathcal{C}^{te}$.

3.2 Multi-Level Adaptation (MLA)

Recent training-based CZSL methods predominantly adopt a multi-path paradigm, which fine-tunes a frozen CLIP model while jointly modeling compositional and primitive-level predictions. Compared to single-path (composition-level) or dual-path (primitive-level) approaches, multi-path models generally achieve superior performance by better capturing compositional structure and semantic dependencies. In this work, we build our MLDA upon a trained multi-path CZSL model. Notably, our approach operates in a plug-and-play manner during inference. Moreover, although developed under the multi-path setting, our method is not limited to this paradigm and can be naturally adjusted to single-path or dual-path CZSL models, demonstrating its flexibility across different choices of $\mathcal{M}^{tr}$ as the base model.

Multi-level Prototype Evolution. Given a trained base model $\mathcal{M}^{tr}$, we first extract visual and textual features for the test samples separately for the composition, attribute, and object branches. Formally, for an input test image $x_t \subset X_{test}$ at time t and a composition $c_i \subset \mathcal{C}^{te}$, we have

$$\boldsymbol{x}_t^c, \boldsymbol{x}_t^a, \boldsymbol{x}_t^o = f(x_t), \quad \boldsymbol{y}_i^c, \boldsymbol{y}^a, \boldsymbol{y}^o = g(c_i, a, o), \tag{1}$$

where $f(\cdot)$ and $g(\cdot)$ denote the visual and textual encoders of $\mathcal{M}^{tr}$. $\boldsymbol{x}_t^c, \boldsymbol{y}_i^c \in \mathbb{R}^d$ denote the composition embeddings, while $\boldsymbol{x}_t^a, \boldsymbol{y}^a \in \mathbb{R}^d$ and $\boldsymbol{x}_t^o, \boldsymbol{y}^o \in \mathbb{R}^d$ denote the attribute and object embeddings, respectively. *For simplicity, the indices of primitive are omitted.*

To further improve the quality of these textual representations at test time, we maintain adaptive textual prototypes that are updated online based on their similarity with the corresponding visual embeddings. Specifically, for the i-th composition prototype $\boldsymbol{y}_i^c$, we compute an update weight:

$$k_{t,i} = \sigma(-\theta \cdot \mathrm{sim}(\boldsymbol{x}_t^c, \boldsymbol{y}_i^c)), \tag{2}$$

where σ is the sigmoid function, θ is a temperature hyperparameter, and $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. The prototype is then updated as:

$$\boldsymbol{y}_i^c \leftarrow \frac{\boldsymbol{y}_i^c + k_{t,i}\tilde{\boldsymbol{y}}_i^c}{\|\boldsymbol{y}_i^c + k_{t,i}\tilde{\boldsymbol{y}}_i^c\|}, \tag{3}$$

where $\tilde{\boldsymbol{y}}^c = [\tilde{\boldsymbol{y}}_1^c, .., \tilde{\boldsymbol{y}}_{|\mathcal{C}^{te}|}^c] \in \mathbb{R}^{|\mathcal{C}^{te}| \times d}$ is a set of learnable parameters initialized to zero. This weighted update encourages adaptation toward compositions that differ from those seen during training, while avoiding excessive adjustments

over the test stream. The same procedure is applied to the primitive-level by optimizing $\tilde{\boldsymbol{y}}^a$ and $\tilde{\boldsymbol{y}}^o$ for the attribute and object branches, yielding adaptive primitive prototypes.

Prototype-Based Inference. Given the multi-level prototypes, the prediction for the t-th test image x_t is computed by:

$$p(u_i \mid x_t) = \frac{\exp\left(\mathrm{sim}(\boldsymbol{x}_t^u, \boldsymbol{y}_i^u)/\tau\right)}{\sum_{j=1}^{|\mathcal{U}|} \exp\left(\mathrm{sim}(\boldsymbol{x}_t^u, \boldsymbol{y}_j^u)/\tau\right)}, \tag{4}$$

where $u \in \{c, a, o\}$ denotes the composition, attribute, or object branch, and $\mathcal{U}$ is the corresponding label set ($\mathcal{C}^{te}$, $\mathcal{A}$, or $\mathcal{O}$). τ is CLIP temperature parameter. The final compositional prediction is then given by:

$$p_{\text{MLA}}(c_i \mid x_t) = \gamma p(c_i|x_t) + (1-\gamma)(p(a|x_t) \times p(o|x_t)), \tag{5}$$

where γ balances the contributions of primitive prediction and composition prediction. This allows complementary integration of semantic information across multiple levels.

3.3 Dynamic Semantic Alignment

As discussed before, a key challenge in test-time CZSL is cross-modal semantic asymmetry. Inspired by how humans perceive novel concepts from multiple perspectives, we enrich each test image and composition and model them as two discrete distributions in the CLIP space. A dynamic OT-based alignment is then employed to adaptively bridge visual and compositional semantics.

Distributed Feature Construction. As shown in Fig.3(a), for each test image x_t, we apply standard data augmentation, including random flipping and random resized cropping, to generate multiple visual views $\{x_t^m\}_{m=1}^{M+1}$, where x_t^1 denotes the original image and the remaining M samples correspond to augmented views. In parallel, we query large language models to generate a set of diverse textual descriptions for each candidate composition c_i, yielding N textual variants $\{z_i^n\}_{n=1}^{N}$.

Then the augmented image and compositional descriptions are subsequently fed into the composition-branch encoders of the base model to get the corresponding embedding $\{\boldsymbol{x}_t^m\}_{m=1}^{M+1}$ and $\{\boldsymbol{z}_i^n\}_{n=1}^{N}$. However, not all augmented views contribute equally, as some capture core semantics while others may be less informative or even noisy. Therefore, we adopt an entropy-based weighting strategy to quantify the importance of each view. For visual domain, we compute the importance weight of the m-th views as:

$$\begin{aligned} \alpha_t^m &= \frac{\exp(H(\boldsymbol{x}_t^m))}{\sum_{m'=1}^{M+1} \exp\big(H(\boldsymbol{x}_t^{m'})\big)}, \\ H(\boldsymbol{x}_t^m) &= \textstyle\sum_{i=1}^{|\mathcal{C}^{te}|} p(\overline{\boldsymbol{z}}_i \mid \boldsymbol{x}_t^m) \log p(\overline{\boldsymbol{z}}_i \mid \boldsymbol{x}_t^m), \end{aligned} \tag{6}$$

where $\overline{\boldsymbol{z}}_i = 1/N \sum_{n=1}^{N} \boldsymbol{z}_i^n$ is the average embedding for composition c_i. Similarly, for the textual domain, we calculate the importance of n-th description according to the original image embedding $\boldsymbol{x}_t^1$:

$$\begin{aligned} \beta_i^n &= \frac{\exp(H(\boldsymbol{z}_i^n))}{\sum_{n'=1}^{N} \exp(H(\boldsymbol{z}_i^n))}, \\ H(\boldsymbol{z}_i^n) &= p(\boldsymbol{z}_i^n \mid \boldsymbol{x}_t^1) \log p(\boldsymbol{z}_i^n \mid \boldsymbol{x}_t^1) \\ &\quad + \textstyle\sum_{j=1,\, j\neq i}^{|\mathcal{C}^{te}|} p(\overline{\boldsymbol{z}}_j \mid \boldsymbol{x}_t^1) \log p(\overline{\boldsymbol{z}}_j \mid \boldsymbol{x}_t^1), \end{aligned} \tag{7}$$

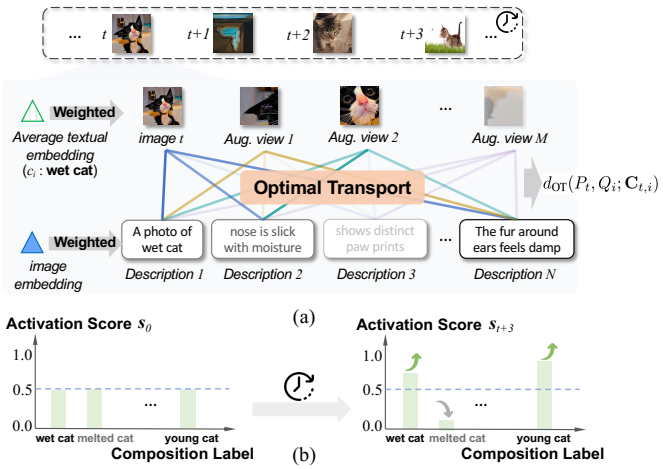


Figure 3: (a) Illustration of our *Dynamic OT Alignment*, where the composition label `wet cat` is activated; (b) Illustration of the *Activation Score* update, where the initial activation score vector $\boldsymbol{s}_0$ is uniformly initialized to 0.5.

This strategy assigns higher weights to views that yield more confident (i.e., lower-entropy) predictions, thereby prioritizing contextually significant views. Based on these, we further model the resulting visual and textual features as two discrete distributions in the CLIP embedding space:

$$P_t = \sum_{m=1}^{M+1} \alpha_t^m \delta_{\boldsymbol{x}_t^m}, Q_i = \sum_{n=1}^{N} \beta_i^n \delta_{\boldsymbol{y}_i^n}, \tag{8}$$

where $\delta(\cdot)$ denotes the Dirac delta function.

Dynamic Optimal Transport (DOT). Given the discrete distribution P_t for the arriving image x_t and Q_i for a candidate composition c_i, effective alignment is critical. Fortunately, Optimal Transport (OT) has been verified as a powerful tool to measure the distance between two distributions [Wang *et al.*, 2023a]. Unlike point-wise similarity metrics, OT is distribution-aware and explicitly accounts for the transport of probability mass between two distributions, thereby facilitating richer interactions across modalities. Concretely, the OT distance between P_t and Q_i is formulated as:

$$\begin{aligned} d_{\text{OT}}(P_t, Q_i; \mathbf{C}_{t,i}) &= \min_{\mathbf{T}_{t,i}} \langle \mathbf{T}_{t,i}, \mathbf{C}_{t,i} \rangle - \epsilon H(\mathbf{T}_{t,i}), \\ \text{s.t.} \quad \mathbf{T}_{t,i} \mathbb{1}^N &= \boldsymbol{\alpha}_t, \quad \mathbf{T}_{t,i}^\top \mathbb{1}^{M+1} = \boldsymbol{\beta}_i, \end{aligned} \tag{9}$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius dot-product and $\mathbf{T}_{t,i} \in \mathbb{R}^{(M+1)\times N}$ is the transport plan, which indicates the transport probability from each visual augmentation in P_t to each composition description in Q_i. $\mathbf{C}_{t,i} \in \mathbb{R}^{(M+1)\times N}$ is the cost matrix, whose elements represent the transport cost computed using cosine distance. $\epsilon > 0$ is a coefficient. $\boldsymbol{\alpha}_t = [\alpha_t^1, ..., \alpha_t^{M+1}]^\top$ and $\boldsymbol{\beta}_t = [\beta_t^1, ..., \beta_t^N]^\top$ are probability vector summing to 1 (seen in Eq.6-7) and are dynamically updated based on the current cross-modal correlations during testing. This formulation enables dynamic, fine-grained cross-modal alignment between visual evidence and compositional semantics. Consequently, the prediction probability of image

x_t can be expressed as:

$$p_{\text{DOT}}(c_i|x_t) = \frac{\exp(-d_{\text{OT}}(P_t, Q_i; \mathbf{C}_{t,i}))}{\sum_{i'=1}^{|\mathcal{C}^{te}|} \exp(-d_{\text{OT}}(P_t, Q_{i'}; \mathbf{C}_{t,i'}))}, \quad (10)$$

Adaptive Activation Scores (AAS). During test-time adaptation, not all compositions require equal attention due to distribution shifts and the partial infeasibility of the composition space. To address this, we assign each candidate composition an adaptive activation score, learned online to reflect its relevance under the current test distribution. Specifically, we define the activation scores at time t as:

$$\boldsymbol{s}_t = [s_{t,1}, ..., s_{t,|\mathcal{C}^{te}|}]^\top, s_{t,i} \in [0, 1] \quad (11)$$

As test images arrive sequentially, the activation scores in MLDA are dynamically updated according to the current test distribution. As illustrated in Fig. 3(b), the score of infeasible compositions, such as melted cat, decreases over time. On the one hand, these activation scores act as a gating mechanism for OT alignment, where a threshold τ selects an active subset $\mathcal{C}^{act}$ of relevant compositions, leading to more efficient and accurate alignment. On the other hand, $\boldsymbol{s}_t$ are used to reweight the aggregated logits in Sec.3.4, allowing more reliable compositions to exert greater influence on the final prediction, thereby promoting robust and discriminative test-time decision making under distribution shifts.

3.4 Optimization and Prediction

At test time, the final prediction is obtained by aggregating the outputs of the multi-level alignment (MLA) module and the dynamic optimal transport (DOT) module. Specifically, for composition c_i at time t, the aggregated logit is computed as:

$$p_{\text{agg}}(c_i|x_t) = (p_{\text{MLA}}(c_i|x_t) + \lambda \cdot p_{\text{DOT}}(c_i|x_t)) \cdot s_{t,i}, \quad (12)$$

where λ denotes the weighting coefficient of DOT module.

To enable online adaptation without access to ground-truth labels, we optimize the model using prediction entropy minimization:

$$\mathcal{L}_{ent}(x_t) = -\sum_{i=1}^{|\mathcal{C}^{te}|} p_{\text{agg}}(c_i \mid x_t) \log p_{\text{agg}}(c_i \mid x_t). \quad (13)$$

where only $\tilde{\boldsymbol{y}}^c, \tilde{\boldsymbol{y}}^a, \tilde{\boldsymbol{y}}^o$, and $\boldsymbol{s}_t$ are learnable parameters. Minimizing prediction entropy promotes confident and consistent outputs, facilitating effective adaptation under test-time distribution shifts.

4 Experiments

4.1 Experimental Settings

Datasets. The proposed MLDA is evaluated on three widely used CZSL benchmark datasets, namely UT-Zappos [Yu and Grauman, 2014], MIT-States [Isola *et al.*, 2015], and C-GQA [Naeem *et al.*, 2021]. UT-Zappos is a fine-grained dataset focusing on footwear, consisting of 16 attributes and 12 object categories. MIT-States contains images from real-world scenarios, where each object is associated with a particular state. It comprises 53,753 images spanning 115 attributes and 245 objects. C-GQA is derived from the Stanford GQA dataset [Hudson and Manning, 2019] and includes over 9,500 compositions formed from 413 attributes and 674 objects, making it the largest benchmark in terms of compositional pairs for CZSL.

Metrics. Following the standard evaluation protocol adopted in prior works [Lu *et al.*, 2023], we assess performance under both closed-world and open-world settings using four evaluation metrics. Specifically, **Seen** denotes the best accuracy on seen compositions when the calibration bias is set to is $-\infty$, while **Unseen** measures the accuracy on unseen compositions when the bias is $+\infty$. To capture the overall performance across seen and unseen pairs, we further report the area under the seen–unseen accuracy curve (**AUC**), obtained by varying the calibration bias from $-\infty$ to $+\infty$, and identify the point that yields the highest harmonic mean (**HM**) between seen and unseen accuracy. Notably, **HM** and **AUC** serve as the primary criteria for comprehensive model evaluation.

Implementation Details. We implement MLDA on the trained base model provided in TsCA [Lu *et al.*, 2023], which adopts the CLIP ViT-L/14 architecture. All experiments are conducted in the PyTorch framework using a single NVIDIA H200 GPU. The value of λ is set to 0.5, 0.35, and 0.3 for the UT-Zappos, MIT-States, and C-GQA datasets, respectively. For augmentation, we apply random resized cropping and horizontal flipping to images, and query Qwen2.5-7B-Instruct to generate diverse compositional descriptions. We set the number of image augmentations to $M = 15$ for each image and the number of compositional descriptions to $N = 15$ for each composition.

4.2 Comparison with State-of-the-Arts

We compare our method with recent state-of-the-art approaches using the same CLIP ViT-L/14 backbone in Tabs. 1–2. For fairness, we also report reproduced results of TMOCAT [Yan and Feng, 2025] using the same base model. In the closed-world setting, MLDA consistently outperforms prior SOTA methods across all datasets. It improves AUC by 1.8% on UT-Zappos, 0.3% on MIT-States, and 0.4% on C-GQA, while achieving corresponding HM gains of 1.9%, 0.2%, and 0.5%. The more pronounced improvement on the fine-grained UT-Zappos dataset highlights MLDA's effectiveness in precise cross-modal alignment. In the open-world setting, our method attains the highest AUC across all datasets, with scores of 45.3%, 9.5%, and 4.9%. Compared with TMOCAT, MLDA demonstrates improved robustness, validating the benefit of accumulating multi-level prototype knowledge for handling noisy test image and unseen compositions.

4.3 Ablation Study

Main Component Analysis. In Tab.3, we assess the effectiveness of each components of MLDA. We have the following key observations: (1) Introducing any individual module (rows 2–4) consistently improves most metrics compared to the base model. (2) The proposed modules are complementary and mutually reinforcing. Specifically, MLA focuses on accumulating multi-level global historical knowledge, while DOT emphasizes fine-grained semantic alignment for the current test sample. In addition, AAS reduces the computational space of DOT to improve efficiency, while also reweighting the final predictions of other modules. (3) By combining the strengths of all components, the complete model achieves the best overall performance across all metrics.

Methods	UT-Zappos				MIT-States				C-GQA			
	AUC	HM	Seen	Unseen	AUC	HM	Seen	Unseen	AUC	HM	Seen	Unseen
CLIP [Radford *et al.*, 2021]	5.0	15.6	15.8	49.1	11.0	26.1	30.2	46.0	1.4	8.6	7.5	25.0
CoOp [Zhou *et al.*, 2022]	18.8	34.6	52.1	49.3	13.5	29.8	34.4	47.6	4.4	17.1	20.5	26.8
Co-CGE [Mancini *et al.*, 2022]	36.3	49.7	63.4	71.3	17.0	33.1	46.7	45.9	5.7	18.9	34.1	21.2
CSP [Nayak *et al.*, 2023]	33.0	46.6	64.2	66.2	19.4	36.3	46.6	49.9	6.2	20.5	28.8	26.8
DFSP [Lu *et al.*, 2023]	36.0	47.2	66.7	71.7	20.6	37.3	46.9	52.0	10.5	27.1	38.2	32.0
GIPCOL [Xu *et al.*, 2024]	36.2	48.8	65.0	68.5	19.9	36.6	48.5	49.6	7.1	22.5	31.9	28.4
Troika [Huang *et al.*, 2024]	41.7	54.6	66.8	73.8	22.1	39.3	49.0	53.0	12.4	29.4	41.0	35.7
CDS-CZSL [Li *et al.*, 2024b]	39.5	52.7	63.9	74.8	22.4	39.2	50.3	52.9	11.1	28.1	38.3	34.2
PLID [Bao *et al.*, 2023]	38.7	52.4	67.3	68.8	22.1	39.0	49.7	52.4	11.0	27.9	38.8	33.0
IMAX [Jiang *et al.*, 2024]	40.6	54.2	69.3	70.7	21.9	39.1	48.7	53.8	12.8	29.8	39.7	35.8
Retri-Aug [Jing *et al.*, 2024]	44.5	56.5	69.4	72.8	22.5	39.2	50.0	53.3	14.4	32.0	45.6	36.0
TsCA [Li *et al.*, 2024a]	46.1	58.5	68.7	75.8	23.0	39.9	51.2	52.9	15.2	33.1	43.8	38.9
CLUSPRO [Qu *et al.*, 2025]	46.6	58.5	70.7	**76.0**	23.8	40.7	**52.1**	54.0	14.9	32.8	44.3	37.8
TOMCAT [Yan and Feng, 2025]	48.3	60.2	74.5	72.8	22.6	39.5	50.3	53.0	16.0	34.0	**45.3**	40.1
TOMCAT†	47.6	60.9	70.5	74.3	23.2	40.1	50.6	53.5	15.7	33.8	44.6	39.5
MLDA (Ours)	**50.1**	**62.8**	**77.7**	75.5	**24.1**	**40.9**	51.9	**54.9**	**16.4**	**34.5**	45.1	**41.2**

Table 1: CZSL comparisons in the closed-world setting. The best results are in **bold**. The second-best results are in underline. (†) denotes results reproduced using our implementation of the base model.

Methods	UT-Zappos				MIT-States				C-GQA			
	AUC	HM	Seen	Unseen	AUC	HM	Seen	Unseen	AUC	HM	Seen	Unseen
CLIP [Radford *et al.*, 2021]	2.2	11.2	15.7	20.6	3.0	12.8	30.1	14.3	0.3	4.0	7.5	4.6
CoOp [Zhou *et al.*, 2022]	13.2	28.9	52.1	31.5	2.8	12.3	34.6	9.3	0.7	5.5	21.0	4.6
Co-CGE [Mancini *et al.*, 2022]	28.4	45.3	59.9	56.2	5.6	17.7	38.1	20.0	0.9	5.3	33.2	3.9
CSP [Nayak *et al.*, 2023]	22.7	38.9	64.1	44.1	5.7	17.4	46.3	15.7	1.2	6.9	28.7	5.2
DFSP [Lu *et al.*, 2023]	30.3	44.0	66.8	60.0	6.8	19.3	47.5	18.5	2.4	10.4	38.3	7.2
GIPCOL [Xu *et al.*, 2024]	23.5	40.1	65.0	45.0	6.3	17.9	48.5	16.0	1.3	7.3	31.6	5.5
Troika [Huang *et al.*, 2024]	33.0	47.8	66.4	61.2	7.2	20.1	48.8	18.7	2.7	10.9	40.8	7.9
CDS-CZSL [Li *et al.*, 2024b]	32.3	48.2	64.7	61.3	8.5	22.1	49.4	21.8	2.7	11.6	37.6	8.2
PLID [Bao *et al.*, 2023]	30.8	46.6	67.6	55.5	7.3	20.4	49.1	18.7	2.5	10.6	39.1	7.5
IMAX [Jiang *et al.*, 2024]	32.3	47.5	68.4	57.3	7.6	21.4	50.2	18.6	2.6	11.2	38.7	7.9
Retri-Aug [Jing *et al.*, 2024]	33.3	47.9	69.4	59.4	8.2	21.8	49.9	20.1	4.4	14.6	44.6	11.8
TsCA [Li *et al.*, 2024a]	37.1	52.2	63.4	69.8	8.7	22.3	50.8	21.7	4.5	14.7	44.3	11.4
CLUSPRO [Qu *et al.*, 2025]	39.5	54.1	71.0	**66.2**	9.3	23.0	51.2	22.1	3.0	11.6	41.6	8.3
TOMCAT [Yan and Feng, 2025]	43.7	57.9	74.1	65.8	8.2	21.7	49.2	21.0	4.2	14.2	**45.1**	10.6
TOMCAT†	42.1	56.7	73.8	63.3	8.8	22.2	51.3	21.6	4.5	14.8	44.5	11.3
MLDA (Ours)	**45.3**	**58.9**	**77.4**	64.8	**9.5**	**23.2**	**51.9**	**22.6**	**4.9**	**15.6**	44.8	**11.9**

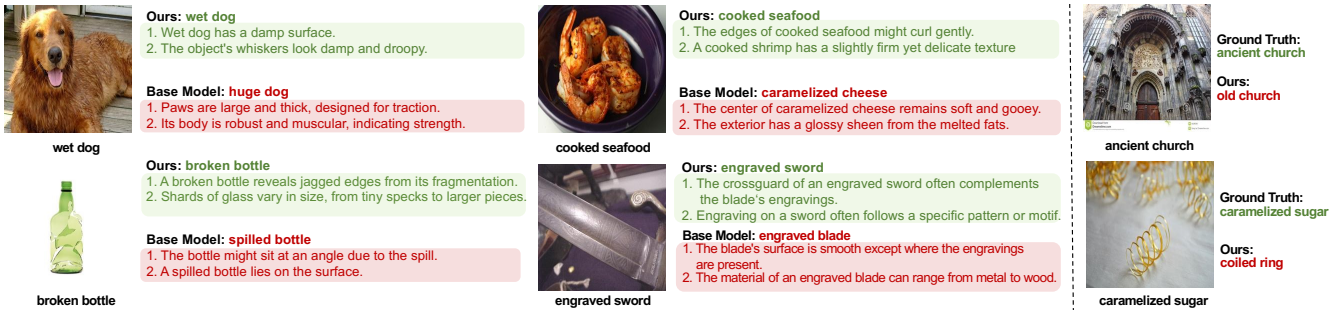
Table 2: CZSL comparisons in the open-world setting. The best results are in **bold**. The second best results are in underline. (†) denotes results reproduced using our implementation of the base model.

Ablation on Base Model. We also evaluate MLDA on different trained base models, with results reported in Tab.4. The results indicate that our method consistently yields significant performance gains across all base models, demonstrating its flexibility and robustness.

Ablation on Multi-Level Prototype. In Tab.5, we evaluate the impact of each branch in the multi-level adaptation module. Introducing either attribute- or object-level prototypes results in performance gains. Notably, retaining both composition- and primitive-level prototypes achieves the best results, highlighting the importance of comprehensive knowledge accumulation during test-time adaptation.

Module			UT-Zappos		MIT-States	
MLA	DOT	AAS	AUC	HM	AUC	HM
			46.1	58.5	23.0	39.9
✓			48.7	61.4	23.1	39.4
	✓		47.1	60.5	23.3	39.9
		✓	46.6	59.3	23.0	39.3
✓	✓		49.1	61.8	23.9	40.8
✓		✓	49.6	62.1	23.2	39.6
	✓	✓	47.3	61.5	23.3	40.0
✓	✓	✓	**50.1**	**62.8**	**24.1**	**40.9**

Table 3: Ablation results for each component on UT-Zappos and MIT-States datasets. MLA means the Multi-Level Adaptation in Sec.3.2. DOT and AAS denote Dynamic Optimal Transport and Adaptive Activation Scores in Sec.3.3



(a) Comparison with base model (b) Failure cases

Figure 4: Qualitative results on MIT-States dataset. (a) For each sample, we show the input image together with its ground-truth composition. We report the predictions from MLDA and the base model, along with the corresponding compositional descriptions. Correct and incorrect compositions are marked in green and red, respectively. (b) Representative failure cases are shown.

Module	UT-Zappos				MIT-States			
	AUC	HM	Seen	Unseen	AUC	HM	Seen	Unseen
CLIP	2.2	11.2	15.7	20.6	3.0	12.8	30.1	14.3
CLIP + Ours	9.4	24.4	49.0	26.5	14.7	30.8	51.2	35.4
CSP	33.0	46.6	64.2	66.2	19.4	36.3	46.6	49.9
CSP + Ours	36.7	51.5	65.8	64.6	21.1	37.9	47.0	53.4
Troika	41.7	54.6	66.8	73.8	22.1	39.3	49.0	53.0
Troika + Ours	45.1	59.5	73.5	67.2	23.5	40.4	50.8	54.5

Table 4: Ablation study of different base models on UT-Zappos and MIT-States datasets in the closed-world setting.

Branch			Closed-world		Open-world	
$\mathcal{C}$	$\mathcal{A}$	$\mathcal{O}$	AUC	HM	AUC	HM
✓			49.3	61.9	43.9	57.5
	✓	✓	48.3	62.0	43.4	58.0
✓	✓		49.8	62.5	45.1	58.6
✓		✓	49.5	62.0	44.0	58.2
✓	✓	✓	**50.1**	**62.8**	**45.3**	**58.9**

Table 5: Ablation on the multi-level prototype on UT-Zappos. $\mathcal{C}$, $\mathcal{A}$ and $\mathcal{O}$ denote composition-, attribute-, and object-level prototypes.

Impacts of the Number of Augmentation views. We analyze how the number of augmented views in both modalities affects performance in Fig. 5. The results show that increasing the number of image augmentations or compositional descriptions consistently improves performance, as richer visual regions and descriptive sentences enhance cross-modal alignment. However, the gains gradually saturate beyond a certain point. Therefore, we set $M = N = 15$ by default.

4.4 Qualitative Analysis

Fig. 4 presents several top-1 prediction examples, where our MLDA shows consistently better performance than the base model. This can be attributed to two main factors. First, multi-level prototype adaptation enables the model to accumulate more comprehensive semantic knowledge across different levels, facilitating effective adaptation to the new distribution. Second, by enriching compositional labels with generated textual descriptions, MLDA is able to better capture the subtle semantic differences between similar compositions. For instance, `wet dog` is characterized by *a damp surface*, whereas `huge dog` is distinguished by attributes such as *large and thick paws*. This enhanced semantic representation allows for finer-grained cross-modal alignment.

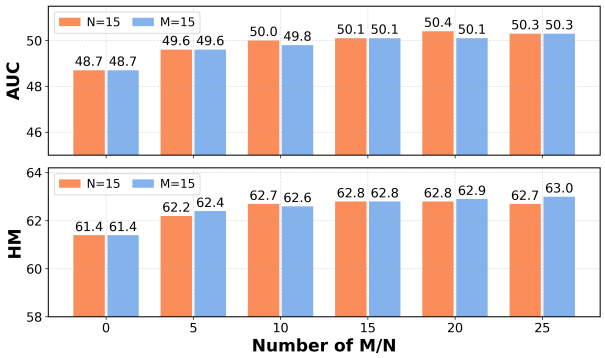


Figure 5: Ablation on the number of text descriptions N and image augmentations M on the UT-Zappos dataset.

Simultaneously, we present some representative failure cases. One type of failure arises from semantic ambiguity, where compositions such as `ancient church` and `old church` are visually indistinguishable in the given images. Another failure mode is caused by incorrect annotations, for example, an image depicting `coiled ring` that is mistakenly labeled as `caramelized sugar`.

5 Conclusion

In this paper, we present MLDA, a novel test-time adaptation framework designed to address label distribution shifts in CZSL. Rather than relying solely on composition-level knowledge, MLDA also leverages informative primitive-level information from historical unlabeled data to provide complementary context. Meanwhile, we model the test image and candidate compositions as discrete distributions to enable fine-grained semantic alignment. Finally, our activation scores dynamically calibrate the final predictions. Extensive experiments and ablation studies demonstrate the effectiveness and flexibility of our approach.

Acknowledgments

This research was supported by funding from the Hong Kong RGC General Research Fund (Grant Nos. 15221123, 15216424, and 15211525), and the Hong Kong PolyU Internal Research Fund (Grant Nos. P0058468 and P0056171).

References

[Bao *et al.*, 2023] Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. Prompting language-informed distribution for compositional zero-shot learning. *arXiv preprint arXiv:2305.14428*, 2023.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Feng *et al.*, 2023] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023.

[Huang *et al.*, 2024] Siteng Huang, Biao Gong, Yutong Feng, Min Zhang, Yiliang Lv, and Donglin Wang. Troika: Multi-path cross-modal traction for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24005–24014, 2024.

[Hudson and Manning, 2019] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[Isola *et al.*, 2015] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015.

[Jiang and Zhang, 2024] Chenyi Jiang and Haofeng Zhang. Revealing the proximate long-tail distribution in compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2498–2506, 2024.

[Jiang *et al.*, 2024] Chenyi Jiang, Shidong Wang, Yang Long, Zechao Li, Haofeng Zhang, and Ling Shao. Imaginary-connected embedding in complex space for unseen attribute-object discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Jing *et al.*, 2024] Chenchen Jing, Yukun Li, Hao Chen, and Chunhua Shen. Retrieval-augmented primitive representations for compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2652–2660, 2024.

[Karmanov *et al.*, 2024] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024.

[Khurana *et al.*, 2021] Ansh Khurana, Sujoy Paul, Piyush Rai, Soma Biswas, and Gaurav Aggarwal. Sita: Single image test-time adaptation. *arXiv preprint arXiv:2112.02355*, 2021.

[Lake *et al.*, 2017] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[Li *et al.*, 2022] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. 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, 2022.

[Li *et al.*, 2023] Yun Li, Zhe Liu, Saurav Jha, and Lina Yao. Distilled reverse attention network for open-world compositional zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1782–1791, 2023.

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

[Li *et al.*, 2024b] Yun Li, Zhe Liu, Hang Chen, and Lina Yao. Context-based and diversity-driven specificity in compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17037–17046, 2024.

[Lu *et al.*, 2023] Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23560–23569, 2023.

[Mancini *et al.*, 2021] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5222–5230, 2021.

[Mancini *et al.*, 2022] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on pattern analysis and machine intelligence*, 46(3):1545–1560, 2022.

[Misra *et al.*, 2017] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017.

[Nado *et al.*, 2020] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

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

[Nayak *et al.*, 2023] Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. Learning to compose soft prompts for compositional zero-shot learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[Qu *et al.*, 2025] Hongyu Qu, Jianan Wei, Xiangbo Shu, and Wenguan Wang. Learning clustering-based prototypes for compositional zero-shot learning. *arXiv preprint arXiv:2502.06501*, 2025.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Rao *et al.*, 2025] Zhijie Rao, Jingcai Guo, Miaoge Li, and Yang Chen. Exploring transferable homogeneous groups for compositional zero-shot learning. *arXiv preprint arXiv:2501.10695*, 2025.

[Saini *et al.*, 2022] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667, 2022.

[Shu *et al.*, 2022] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[Tanwisuth *et al.*, 2023] Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *International Conference on Machine Learning*, pages 33816–33832. PMLR, 2023.

[Villani and others, 2008] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

[Wang *et al.*, 2020] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[Wang *et al.*, 2023a] Dongsheng Wang, Miaoge Li, Xinyang Liu, MingSheng Xu, Bo Chen, and Hanwang Zhang. Tuning multi-mode token-level prompt alignment across modalities. *Advances in Neural Information Processing Systems*, 36:52792–52810, 2023.

[Wang *et al.*, 2023b] Henan Wang, Muli Yang, Kun Wei, and Cheng Deng. Hierarchical prompt learning for compositional zero-shot recognition. In *IJCAI*, volume 1, page 3, 2023.

[Wang *et al.*, 2024] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[Xiao *et al.*, 2025] Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees GM Snoek. Dynaprompt: Dynamic test-time prompt tuning. *arXiv preprint arXiv:2501.16404*, 2025.

[Xu *et al.*, 2024] Guangyue Xu, Joyce Chai, and Parisa Kordjamshidi. 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, 2024.

[Yan and Feng, 2025] Xudong Yan and Songhe Feng. Tomcat: Test-time comprehensive knowledge accumulation for compositional zero-shot learning. *arXiv preprint arXiv:2510.20162*, 2025.

[Yu and Grauman, 2014] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 192–199, 2014.

[Zhang *et al.*, 2022] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

[Zhou and Ma, 2024] Sicong Zhou and Jianhui Ma. Test-time prompt tuning for compositional zero-shot learning. In *2024 3rd International Conference on Electronics and Information Technology (EIT)*, pages 745–749. IEEE, 2024.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.