Revealing the Proximate Long-Tail Distribution in Compositional Zero-Shot Learning

Chenyi Jiang, Haofeng Zhang *

School of Computer Science and Engineering, Nanjing University of Science and Technology, China
{jiangchenyi, zhanghf}@njust.edu.cn,

Abstract

Compositional Zero-Shot Learning (CZSL) aims to transfer knowledge from seen state-object pairs to novel unseen pairs. In this process, visual bias caused by the diverse interrelationship of state-object combinations blurs their visual features, hindering the learning of distinguishable class prototypes. Prevailing methods concentrate on disentangling states and objects directly from visual features, disregarding potential enhancements that could arise from a data viewpoint. Experimentally, we unveil the results caused by the above problem closely approximate the long-tailed distribution. As a solution, we transform CZSL into a proximate class imbalance problem. We mathematically deduce the role of class prior within the long-tailed distribution in CZSL. Building upon this insight, we incorporate visual bias caused by compositions into the classifier's training and inference by estimating it as a proximate class prior. This enhancement encourages the classifier to acquire more discernible class prototypes for each composition, thereby achieving more balanced predictions. Experimental results demonstrate that our approach elevates the model's performance to the state-of-the-art level, without introducing additional parameters. Our code is available at https://github.com/LanchJL/ProLT-CZSL.

Introduction

Objects in the world often exhibit diverse states of existence; an *apple* can be *sliced* or *unripe*, while a *building* can be *ancient* or *huge*. Humans have the ability to recognize the composition of the unseen based on their knowledge of seen elements. Even if people have never seen a *green apple* before, they can infer the characteristics of a *green apple* from a *red apple* and a *green lemon*. To empower the machine with this capability, previous work (**??**) propose Compositional Zero-Shot Learning (CZSL), a task aims to identify unseen compositions from seen state-object compositions.

However, the combination of state-objects creates a visual bias for a attribute (state or object) in it, hindering the learning of distinguishable class prototypes. In the face of above challenge, early approaches in the domain of CZSL can be categorized into two distinct methods. The first method utilized two independent classifiers to categorize states and objects (**????**). The second method involved training a common embedding space where semantic and visual features could be projected to reduce the distance between them (**???**). Commonly, these studies concentrate on improving the structure of classifiers and investigating alternative architectures. However, minimal research has been conducted considering the problem in terms of data distribution.

We analyze the prior and posterior probabilities associated with attributes (states or objects) and compositions to determine a more suitable solution. Fig. 1 illustrates that the class prior follows a distinct trend differing from the posterior probabilities. For instance, even though the model is trained on a comparable number of samples, it demonstrates a low probability of predicting the object labeled as *O5*. This issue also extends to making inferences about compositions, which reminds us of the long-tail distribution or class imbalance (**???**).

We consider that certain samples are infected by the intricate interplay between objects and states within compositions (**?**), leading to significant bias from the ideal class prototype. Consequently, these samples with large visual bias make it difficult for the classifier to fit their intrinsic patterns, results in the inability to form effective classification boundaries. In contrast to class imbalance, we refer to this phenomenon as *'attribute imbalance'* below. The recent methods for CZSL (**??**) synchronize the prediction of visual features to states and objects with the prediction of compositions in the common embedding space, which works as a model ensemble approach. While this design addresses the capability to categorize some classes, the non-interaction among the independent classifiers may lead to incomplete mutual compensation due to potential information gaps.

The identified shortcomings prompted a redesign of the model using the model ensemble approach. Building on the success of logit adjustment in addressing long-tail learning (**?**), this study treats attribute imbalance information as special prior knowledge (In the following we denote by *'attribute prior'*) that approximates the class prior. This attribute prior is derived from the estimation of available samples by two independent classifiers for states and objects. In other words, we construct this prior by modelling the visual bias of states and objects from samples. During the training phase, we incorporate it through logit adjustment into the common embedding space. This approach enables

*Corresponding author

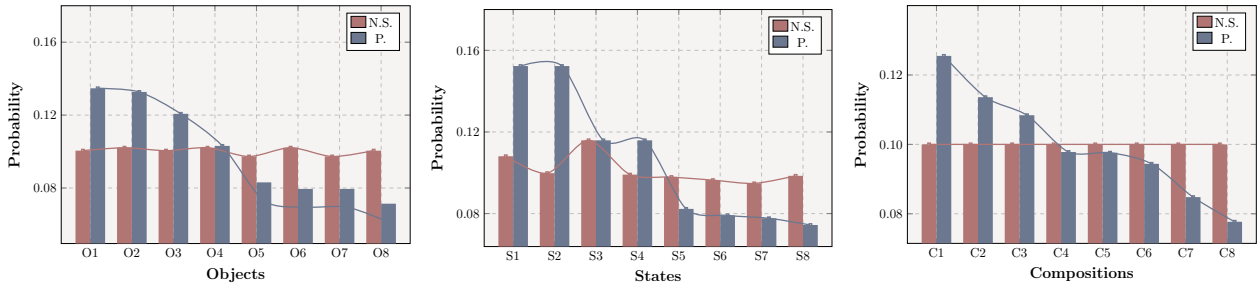

Figure 1: An example of prior and posterior probabilities (predicted by model) of same classes in MIT-States (**?**). N.S. represents the prior probability calculated from the number of samples in each class. P. denotes the average value of posterior probabilities indicating the likelihood of the sample belonging to its class, which is predicted by a MLP with ResNet-18 (**?**) as backbone (same settings as $\mathcal{C}_o, \mathcal{C}_s$ and $\mathcal{C}_y$ in Implementation Details and trained via vanilla cross-entropy loss). The classes are selected on the basis of the closest sample size, and shows the results for the object (left), state (middle), and composition (right) classes. All data is simply normalized for ease of presentation.

the production of balanced posterior probabilities regarding the poorly-classified classes in Fig. 1, thereby preventing each independent classifier from ineffectively reinforcing the ability to classify the well-classified classes.

Specifically, we reconstructed the CZSL problem from the perspective of mutual information and adjusted the posterior values predicted by the model from the perspective of maximizing mutual information. In addition, we generalize the above attribute prior to the unseen class in order to optimize the lower bound of seen-unseen balanced accuracy (**?**) obtained by **?**. We refer to this method as the logit adjustment for **Pro**ximate **L**ong-**T**ail Distribution (ProLT) in CZSL. Unlike previous methods, ProLT does not necessitate introducing additional parameters, yet it significantly enhances the overall CZSL model performance. Our contributions are summarized as follows:

- In our study, we conduct an analysis of the data distribution in CZSL. We translate the visual bias in compositions into an attribute imbalance and thereby generalize CZSL to a proximate long-tail learning problem.
- Our analysis involves a mathematical examination of both the training and inference phases of the model. This enables us to adapt the model's posterior probability based on the attribute prior.
- Our model enhances the prediction of relationships in compositions without the need for introducing additional parameters. Experimental results on three benchmark datasets demonstrate the effectiveness of our approach.

Related Work

Compositional Zero-Shot Learning (CZSL): Zero-Shot Learning (ZSL) transfers knowledge from seen classes to unseen ones by leveraging attributes (**?????**). CZSL (**????**) builds upon this foundation by incorporating the notion of composition learning (**?**), with its extension primarily relying on the shared semantics of state and object within the composition of both seen and unseen classes.

Initial CZSL methodologies directly classify states and objects, effectively converting the task into a conventional supervised assignment (**?????**). However, the fusion of state-object pairs led to visual bias in both elements, impeding the acquisition of discernible class prototypes. Numerous subsequent strategies utilize visual-semantic alignment within a common embedding space (**???**) to grasp the entwined nature of objects and states within compositions. However, this technique is susceptible to domain shift challenges. Recent methodologies typically amalgamate these two models, creating a framework of model ensembles. For instance, **?** enhances the model's adaptability to unseen classes by disentangling visual features and subsequently reconstituting them for novel classes. Meanwhile, **?** introduces conditional state generation to address visual alterations arising from object-state combination. ProLT aligns closely with this paradigm, although with a greater emphasis on direct inquiries into visual bias attributes.

Long-Tailed Classification: Numerous studies address the issue of imbalanced class distributions, with one prominent approach being posterior modification methods (**?????**). Within ZSL, **?** regards it as an imbalanced challenge involving seen and unseen classes, and then applies regulatory techniques based on logit adjustment. However, this approach does not readily extend to the issue of attribute imbalance in our context. **?** considers the presence of visual bias in samples re-weighting within the optimization process, but its localization-based weighting strategy ignores the differences between classes. In this study, we introduce advanced logit adjustment strategies theoretically, aiming to enhance the equilibrium of predictions between various classes.

Methodology

Task Definition

Considering the two disjoint sets $\mathcal{Y}^S$ and $\mathcal{Y}^U$, *i.e.*, $\mathcal{Y}^S \cap \mathcal{Y}^U = \emptyset$. CZSL aims to classify sample $\mathbf{x} \in \mathcal{X}$ into a composition $y = (s, o) \in \mathcal{Y}$, where $\mathcal{Y} = \mathcal{Y}^S \cup \mathcal{Y}^U$, and samples from $\mathcal{Y}^U$ are unseen during training. y is composed by state $s \in \mathcal{S}$ and object $o \in \mathcal{O}$, $\mathcal{S}$ and $\mathcal{O}$ are sets of states and objects. Samples from $\mathcal{Y}^S$ and $\mathcal{Y}^U$ share the same objects

o and states s, but their compositions (s, o) are different. Define the visual space $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and d_x is the dimension of the space, $\mathcal{X}$ can be divided into $\mathcal{X}^S$ and $\mathcal{X}^U$ based on whether their samples belong to seen classes. We can define the train set as $\mathcal{D}_{seen} = \{(\mathbf{x}, y)|\mathbf{x} \in \mathcal{X}^S, y \in \mathcal{Y}^S\}$ and an unseen set for evaluation of methods which is $\mathcal{D}_{unseen} = \{(\mathbf{x}, y)|\mathbf{x} \in \mathcal{X}^U, y \in \mathcal{Y}^U\}$. We employ the Generalized ZSL setup defined in **?**, which requires both seen and unseen classes involves in testing.

Empirical Analysis on Model Ensemble

Method	M.E.	Sta.	Obj.	S.	U.	HM
MLP (**?**)	F	27.9	31.8	25.3	24.6	16.4
	T	27.9	32.0	29.8	24.5	17.9
GCN (**?**)	F	27.9	32.0	28.7	25.3	17.2
	T	28.3	33.4	28.9	26.0	18.8
I.C.	-	25.3	24.8	19.3	19.0	12.0

Table 1: The results of methods that incorporate a composition classifier, along with the addition of two classifiers for states and objects on top of them (*i.e.*, model ensemble), are presented. I.C. indicates only two independent classifiers are used. M.E. indicates the utilization of model ensemble in the methods, where F denotes false and T denotes true. The metrics in the table are defined in Evaluation Protocol.

For the problem of approximate long-tailed distributions caused by visual bias in CZSL, ensemble-based methods have demonstrated exceptional performance in CZSL (**??**). Typically, this approach combines the predictions of two models to produce the final prediction. The first model consists of two independent classifiers C_o and C_s for objects and states. The second model is a composition classifier C_y. The process of the model can be viewed as inputting the samples into three classifiers to estimate the posterior probabilities:

$$\begin{aligned} &p(s|\mathbf{x}) = \text{softmax}[C_s(\mathbf{x})], p(o|\mathbf{x}) = \text{softmax}[C_o(\mathbf{x})], \\ &p(y|\mathbf{x}) = \text{softmax}[C_y(\mathbf{x})], \\ &\hat{p}(y|\mathbf{x}) = \delta p(y|\mathbf{x}) + (1-\delta)\left[p(s|\mathbf{x}) + p(o|\mathbf{x})\right], \end{aligned} \tag{1}$$

where $p(s|\mathbf{x}), p(o|\mathbf{x})$ and $p(y|\mathbf{x})$ are posterior probability from classifiers, $\hat{p}(y|\mathbf{x})$ is the final posterior probabilities. δ is a weight factor. $\mathcal{C}_y(\mathbf{x})$ denotes the logits for class y based on sample $\mathbf{x}$, and $\mathcal{C}_s(\mathbf{x}), \mathcal{C}_o(\mathbf{x})$ are similarly defined.

As demonstrated in Tab. 1, augmenting two additional posterior estimates $p(o|\mathbf{x})$ and $p(s|\mathbf{x})$ to $p(y|\mathbf{x})$ can significantly enhance CZSL results. However, only relying solely on $p(o|\mathbf{x})$ and $p(s|\mathbf{x})$ does not enable accurate estimation, this suggests the improvement in results is not due to the introduction of superior classifiers. Consequently, we can deduce the subsequent conjectures: The effectiveness of ensemble-based methods emanates from incorporating $\mathcal{C}_s$ and $\mathcal{C}_o$, aiding in the classification of compositions that encounter a relative disadvantage within $\mathcal{C}_y$. While attribute imbalances vary across states, objects, and compositions, all three elements might not simultaneously experience large visual bias for a particular class. Based on these preliminary studies, we can posit that effective classification of classes with large visual bias within common embedding spaces requires information compensation. In our study, we directly estimate visual bias as compensation described above. Considering that the visual bias generated by state-object combination is difficult to eliminate directly, we try to introduce it as an attribute prior into the training process from the classifier. In the following, we detail this process.

From the Perspective of Mutual Information

Let us first consider the problem from a simple CZSL approach based on common embedding spaces like (**??**). The optimization objective of these methods can be viewed as the maximum likelihood:

$$\text{argmin}_\theta \, \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{seen}}[-logp(y|\mathbf{x})], \tag{2}$$

where $p(y|\mathbf{x})$ is defined in Eq. 1, which denotes the distribution of compositions predicted by the model. θ denotes the model parameters.

Given the characteristics of CZSL, where each sample is associated with two labels, s and o, there is a conditionality between the two in the setup of the dataset, *i.e.*,

$$p(y) = p(s, o) = p(o)p(s|o), \tag{3}$$

$p(y)$ and $p(o)$ denotes class prior of class y and object o, and $p(s|o)$ is conditional class prior of s and o.

Inspired by **?**, we look at the above issues through the perspective of mutual information (**?**), we have:

$$\begin{aligned} &I(Y; X) \approx \mathbb{E}_y D_{KL}[p(y|\mathbf{x})\|p(y)] \\ =&\sum_{\mathbf{x},y} p(\mathbf{x}, y) log\frac{p(y|\mathbf{x})}{p(y)} \\ =&\sum_{\mathbf{x},y} p(\mathbf{x}, y) log\frac{p(y|\mathbf{x})}{p(o)p(s|o)}, \end{aligned} \tag{4}$$

where X and Y are discrete random variables corresponding to $\mathbf{x}$ and y, respectively, and S and O are similarly defined. D_{KL} represents the Kullback-Leibler divergence, while $p(\mathbf{x}, y)$ represents the joint probability of the class y and the visual feature $\mathbf{x}$. Due to the real posterior probability between y and $\mathbf{x}$ is unknown, we use $p(y|\mathbf{x})$ as an approximation. We can interpret the optimization of maximum likelihood as follows, based on the posterior term in Eq. 4,

$$log\frac{p(y|\mathbf{x})}{p(o)p(s|o)} \sim \mathcal{C}_y(\mathbf{x}), \tag{5}$$

which can be transfer to:

$$logp(y|\mathbf{x}) \sim \mathcal{C}_y(\mathbf{x}) + logp(o)p(s|o), \tag{6}$$

here, $\mathcal{C}_y(\mathbf{x})$ represents the logits for class y, defined in Eq. 1, $\sim$ denotes approximately equal. The expression on the right-

hand side is re-normalized using the $\mathrm{softmax}$ function, *i.e.*,

$$\begin{aligned} & -logp(y|\mathbf{x}) \\ & \sim log\left[1+\sum_{o_i\neq o}\sum_{s_j\neq s}\frac{p(o_i)p(s_j|o_i)}{p(o)p(s|o)}e^{\mathcal{C}_{\hat{y}}(\mathbf{x})-\mathcal{C}_y(\mathbf{x})}\right] \\ & \sim log\left[1+\sum_{o_i\neq o}\sum_{s_j\neq s}\left(\frac{p(o_i)p(s_j|o_i)}{p(o)p(s|o)}\right)^{\eta}e^{\mathcal{C}_{\hat{y}}(\mathbf{x})-\mathcal{C}_y(\mathbf{x})}\right], \end{aligned} \tag{7}$$

where $\hat{y} = (s_i, o_i)$, and η is an adjustment factor. Eq. 7 demonstrates that by incorporating the class prior $p(s|o)$ and $p(o)$ for state s and object o, we can optimize the model's mutual information. Consequently, we approach the CZSL problem from the perspective of mutual information.

Estimating the Attribute Prior

The above idea comes from the logits adjustment (**?**) introduced to address class imbalance (**??**), which demonstrate that the inclusion of a class prior enhances the maximization of mutual information, and we generalize it to CZSL task.

As stated in Introduction, we undertake the transformation of CZSL into an approximate long-tailed distribution issue caused by visual bias from state-object combinations. Our argument centers on the proposition that attribute imbalance within CZSL contributes to an approximate form of class imbalance, since visual bias hinders reduces the distinguishability of some of the samples. Therefore, exclusive reliance on the class prior is inadequate. Building upon this rationale, we propose to use the attribute prior to assume the function of the class prior within the long-tailed distribution, serving as an approximation.

We propose incorporating the model's conditional posterior probabilities as an approximation for this scenario. We continue to denote it as the 'prior' due to its function as a prior probability during the training process, despite being computed using posterior probability. Since attribute imbalance cannot be directly quantified from the dataset, we simulate it by utilizing the posterior probability of the additional classifiers, for $\mathbf{x}$ and its corresponding s, o, we have:

$$\hat{p}(s) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[p(s|\mathbf{x})], \hat{p}(o) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[p(o|\mathbf{x})], \tag{8}$$

where $\mathbf{x} \in \mathcal{D}_{seen}$, $p(s|\mathbf{x})$ and $p(o|\mathbf{x})$ are defined in Eq. 1, which are posterior probabilities from $\mathcal{C}_s$ and $\mathcal{C}_o$, we use their predicted expectations for all training samples as a special attribute prior. From this we can replace the class prior in Eq. 7 with following item:

$$k(s,o) = \mathrm{softmax}\left[\sigma(s,o)\hat{p}(s)\hat{p}(o)\right], \tag{9}$$

where $\sigma(s,o)$ is a function used to model the conditional nature of the composition, *i.e.*,

$$\sigma(s,o) = \begin{cases} 1 & (s,o) \in \mathcal{Y}^S \cup \mathcal{Y}^U, \\ 0 & else. \end{cases} \tag{10}$$

From this we obtain the final objective function according to Eq. 7:

$$\mathcal{L}_{cls} = log\left[1+\sum_{o_i\neq o}\sum_{s_j\neq s}\left(\frac{k(s_j,o_i)}{k(s,o)}\right)^{\eta}e^{\mathcal{C}_{\hat{y}}(\mathbf{x})-\mathcal{C}_y(\mathbf{x})}\right]. \tag{11}$$

Logit Adjustment for Inference

Due to the introduction of unseen classes in the inference phase we need to make additional adjustments. CZSL usually measures model performance in terms of $\mathcal{A}^H$ which denotes Harmonic Mean (HM) accuracy:

$$\mathcal{A}^H = 2/(\frac{1}{\mathcal{A}^S} + \frac{1}{\mathcal{A}^U}), \tag{12}$$

where $\mathcal{A}^S$, $\mathcal{A}^U$ denote seen and unseen accuracy. **?** provides a lower bound of HM, below we briefly describe its conclusions. For HM's lower bound we have:

$$\mathcal{A}^H \geq 1/\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\frac{|\mathcal{Y}|p(\mathcal{Y})p(y|y\in\mathcal{Y})}{q(\mathcal{C}_{out}=y|\mathbf{x})p(y|\mathbf{x})}, \tag{13}$$

where $q(\mathcal{C}_{out} = y|\mathbf{x})$ represents the probability of predicting class y using our model. The set $\mathcal{Y}$ can be either $\mathcal{Y}^S$ or $\mathcal{Y}^U$, $p(y|y \in \mathcal{Y})$ represents the conditional class prior, and $|\mathcal{Y}|p(\mathcal{Y})$ can be seen as a hyper-parameter that quantifies the differences between seen and unseen classes. Considering that the gap between the domains of seen and unseen classes in CZSL is not significant, we can simply treat $|\mathcal{Y}|p(\mathcal{Y})$ as an ignorable constant in the following process.

Finding the Bayesian optimum for $\mathcal{A}^H$ is difficult. However, it is possible to maximize its lower bound, which is equal to minimizing the upper bound of its inverse, *i.e.*, the denominator term of Eq. 13 is minimized if:

$$\tilde{y} = \mathrm{argmax}_y\left[\mathcal{C}_y(\mathbf{x}) + \eta log p(y|y\in\mathcal{Y})\right], \tag{14}$$

where η is from Eq. 11, $\tilde{y}$ is the predicted label for sample $\mathbf{x}$. For conditional class prior $p(y|y \in \mathcal{Y}^S)$, which represents the true class frequency when y belong to seen classes. Following Eq. 11, we similarly replace the prior with the attribute prior estimate in Eq. 9 here, which is:

$$p(y|(s,o)\in\mathcal{Y}^S) := k(s,o), \quad (s,o)\in\mathcal{Y}^S, \tag{15}$$

and the attribute prior of unseen classes are not available to the model, we model it here using a combination of the estimation from Importance Sampling (**?**) with the attribute prior from seen samples, which can be denoted as:

$$p(y|y\in\mathcal{Y}^U) := k(s,o) + \frac{\hat{k}_{\mathbf{x}}(s,o)}{\lambda k(s,o)}, \quad (s,o)\in\mathcal{Y}^U, \tag{16}$$

where $\frac{1}{\lambda}$ is a hyper-parameter denotes the distribution of $\mathbf{x}$. The above results are re-transformed into probability distributions in the actual calculation. And $\hat{k}_{\mathbf{x}}(s,o)$ is instance-based conditional posterior probability:

$$\hat{k}_{\mathbf{x}}(s,o) = \mathrm{softmax}\left[\sigma(s,o)p(s|\mathbf{x})p(o|\mathbf{x})\right], \tag{17}$$

the aforementioned setup arises because during testing, we are unable to provide posterior probabilities $\hat{k}_{\mathbf{x}}(s,o)$ from

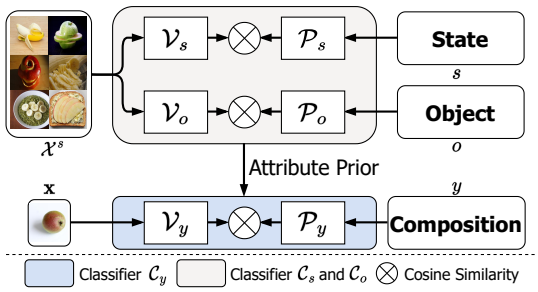

Figure 2: A brief demonstration of ProLT's training stage (detailed in Method Overview). $\mathcal{X}^s$ is the set of seen visual features, we obtain the attribute prior according to Eq. 8.

multiple samples simultaneously. Furthermore, Importance Sampling results in significant variance when the number of samples is insufficient. To address this, we attempt to augment it by leveraging seen attribute prior.

ProLT makes inferences during testing phase based on Eq. 14, our aim is to integrate local information during testing with the prior derived from seen classes, to address the disparities between seen and unseen classes. With Eq. 14 ProLT theoretically achieves the best overall accuracy.

Method Overview

This section provides a concise summary of the aforementioned methods. Our approach, illustrated in Fig. 2, involves training two independent classifiers denoted as $\mathcal{C}_s$ and $\mathcal{C}_o$. These classifiers are implemented using prototype learners, namely $\mathcal{P}_s$ and $\mathcal{P}_o$, and visual embedders $\mathcal{V}_o$, $\mathcal{V}_s$, to determine the prototypes of states and objects, *i.e.*,

$$\mathcal{C}_s(\mathbf{x}) := \frac{cos(\mathcal{V}_s(\mathbf{x}), \mathcal{P}_s(s))}{\tau}, \mathcal{C}_o(\mathbf{x}) := \frac{cos(\mathcal{V}_o(\mathbf{x}), \mathcal{P}_o(o))}{\tau}, \tag{18}$$

where τ is the temperature. These classifiers are trained with vanilla cross-entropy loss:

$$\mathcal{L}_{ic} = log[1 + \sum_{s' \neq s} e^{\mathcal{C}_{s'}(\mathbf{x}) - \mathcal{C}_s(\mathbf{x})}][1 + \sum_{o' \neq o} e^{\mathcal{C}_{o'}(\mathbf{x}) - \mathcal{C}_o(\mathbf{x})}]. \tag{19}$$

Once the classifiers reach a specific training stage, we calculate the attribute prior using Eq. 8, and employ the loss function $\mathcal{L}_{cls}$ from Eq. 11 for training the classifier $\mathcal{C}_y$ for compositions:

$$\mathcal{C}_y(\mathbf{x}) := \frac{cos(\mathcal{V}_y(\mathbf{x}), \mathcal{P}_y(y))}{\tau}, \tag{20}$$

where $\mathcal{P}_y$ is the prototype learner for compositions and $\mathcal{V}_y$ is a visual embedder. After training, the model uses Eq. 14 for inference.

Experiments

Datas

There are numerous recent approaches to compositionality research, and three datasets have been primarily employed for evaluation: MIT-States (**?**), UT-Zappos (**?**), and C-GQA (**?**). We utilized a standardized evaluation dataset for a reasonable comparison with previous methods.

MIT-States presents a considerable challenge, consists of 53,753 images. It comprises 115 state classes, 245 object classes, and 1,962 compositions. In the total compositions, there are 1,262 seen compositions, and 700 compositions remain unseen. **UT-Zappos** is a collection of 50,025 images that focuses on various forms of footwear. It consists of 12 object classes and 16 state classes which is a fine-grained dataset, yielding 116 compositions, of which 83 are seen. **C-GQA** is introduced by **?**, which encompasses a wide variety of real-world common objects. It comprises 413 states, 674 objects, and over 27,000 images, along with more than 9,000 compositions, consisting of 5,592 seen and 1,932 unseen compositions.

Evaluation Protocol

The setting of GZSL (**?**) requires both seen and unseen compositions during testing. We report the **best accuracy** of **seen classes (best seen)**, the **unseen class (best unseen)**, and its **harmonic accuracy (HM)**. In order to measure the performance on attribute learning, we report the **best accuracy** of **states (best sta)** and **objects (best obj)**. Building upon the research of (**?**) and (**?**), we calculate the **Area Under the Curve (AUC)** by comparing the accuracy on seen and unseen compositions with various bias terms.

Implementation Details

Below we present the details of the implementation of ProLT on ResNet-18 (**?**).

Visual Representations and Semantic: In line with prior methods, we employed ResNet-18 pre-trained on ImageNet (**?**) to extract 512-dimensional visual features from the images. For semantic information, we utilized GloVe (**?**) to extract attribute names as 300-dimensional word vectors.

Implementations and Hyper-Parameters: For three prototype learner $\mathcal{P}_s, \mathcal{P}_o$ and $\mathcal{P}_y$ are GloVe connects with two Fully Connected (FC) layers with ReLU (**?**) following the first layer. And the three visual embedders $\mathcal{V}_s, \mathcal{V}_o$, and $\mathcal{V}_y$ are also two FC layers with ReLU and Dropout (**?**). All FCs embed the input features in 512 dimensions and the hidden layer is 1024 dimensions. The overall model is trained using the Adam optimizer (**?**) on NVIDIA GTX 2080Ti GPU, and it is implemented with PyTorch (**?**). We set the learning rate as 5×10^{-4} and the batchsize as 128. We train the $\mathcal{C}_s, \mathcal{C}_o$ and $\mathcal{C}_y$ with an early-stopping strategy, it needs about 400 epochs on MIT-States, 300 epochs on UT-Zappos and 400 epochs on C-GQA. For hyper-parameters, we set τ as $0.1, 0.1, 0.01$, η as $1.0, 1.0, 1.0$ and λ as $50, 10, 100$ for MIT-States, UT-Zappos, and C-GQA, respectively.

Compared with State-of-the-Arts

ProLT is mainly compared with recent methods using fixed ResNet-18 as backbone with the same settings. We also compared ProLT with the CLIP-based approaches (**??**) after using CLIP (**?**) to learn visual and semantic embeddings. The comparison results are shown in Tab. 2.

	Methods	MIT-States						UT-Zappos						C-GQA					
		AUC	HM	S.	U.	Sta.	Obj.	AUC	HM	S.	U.	Sta.	Obj.	AUC	HM	S.	U.	Sta.	Obj.
†	LE+ (**?**)	2.0	10.7	15.0	20.1	23.5	26.3	25.7	41.0	53.0	61.9	41.2	69.2	0.8	6.1	18.1	5.6	-	-
	AttOp (**?**)	1.6	9.9	14.3	17.4	21.1	23.6	25.9	40.8	59.8	54.2	38.9	69.6	0.7	5.9	17.0	5.6	-	-
	TMN (**?**)	2.9	13.0	20.2	20.1	23.3	26.5	29.3	45.0	58.7	60.0	40.8	69.9	1.1	7.5	23.1	6.5	-	-
	SymNet (**?**)	3.0	16.1	24.4	25.2	26.3	28.3	23.9	39.2	53.3	57.9	40.5	71.2	2.1	11.0	26.8	10.3	-	-
	CompCos (**?**)	4.5	16.4	25.3	24.6	27.9	31.8	28.7	43.1	59.8	62.5	44.7	73.5	2.6	12.4	28.1	11.2	-	-
	CGE (**?**)	5.1	17.2	28.7	25.3	27.9	32.0	26.4	41.2	56.8	63.6	45.0	73.9	2.3	11.4	28.1	10.1	-	-
	SCEN (**?**)	5.3	18.4	29.9	25.2	28.2	32.2	32.0	47.8	63.5	63.1	47.3	75.6	2.9	12.4	28.9	12.1	13.6	27.9
	Co-CGE (**?**)	5.1	17.5	27.8	25.2	-	-	29.1	44.1	58.2	63.3	-	-	2.8	12.7	29.3	11.9		
	OADis (**?**)	5.9	18.9	31.1	25.6	28.4	33.2	30.0	44.4	59.5	65.5	46.5	75.5	-	-	-	-	-	-
	DECA (**?**)	5.3	18.2	29.8	25.5	-	-	31.6	46.3	62.7	63.1	-	-	-	-	-	-	-	-
	CANet (**?**)	5.4	17.9	29.0	26.2	30.2	32.6	33.1	47.3	61.0	66.3	48.4	72.6	**3.3**	**14.5**	30.0	13.2	17.5	22.3
	ProLT (Ours)	**6.0**	**19.3**	30.9	26.5	29.5	34.2	**33.4**	**49.3**	62.7	64.0	46.1	74.2	3.2	14.4	32.1	13.7	17.8	32.5
‡	CSP (**?**)	19.4	36.3	46.6	49.9	-	-	33.0	46.6	66.2	64.2	-	-	6.2	20.5	26.8	28.8	-	-
	DFSP (**?**)	20.8	37.7	52.8	47.1	-	-	36.0	47.2	66.7	71.7	-	-	10.5	27.1	38.2	32.0		
	ProLT (Ours)	**21.1**	**38.2**	49.1	51.0	49.8	59.0	**36.1**	**49.4**	66.0	70.1	52.6	79.4	**11.0**	**27.7**	39.5	32.9	24.9	50.1

Table 2: The SoTA comparisons on three datasets. We compare ProLT with others on AUC, best HM, best sta (Sta.), best obj (Obj.), best seen (S.) and best unseen (U.). † denotes ResNet-based methods and ‡ denotes CLIP-based methods. The best AUC and HM for ResNet-based methods and CLIP-based methods are shown in bold.

Method	Prior	AUC	HM	S.	U.	Sta.	Obj.
GCN	N.	28.4	45.0	58.9	60.0	43.4	70.0
	C.P.	29.6	46.2	58.9	61.3	43.9	71.8
	A.P.	32.3	48.4	61.7	62.4	45.9	73.2
FC	N.	32.3	46.8	61.1	64.8	44.1	72.3
	C.P.	32.0	47.9	62.2	62.1	44.2	73.7
	A.P.	33.4	49.3	62.7	64.0	46.1	74.2

Table 3: A comparison of different priors for Eq. 11 and Eq. 14 on UT-Zappos. GCN: GCN is used as the prototype learner, FC: FC layers are used as the prototype learner. N. represents no prior is introduced, using pure ensemble method, C.P. represents class prior from datasets is utilized, and A.P. denotes attribute prior is incorporated.

The results demonstrate that ProLT achieves a new state-of-the-art performance when using ResNet-18 as backbone on the MIT-States, UT-Zappos, and C-GQA. Specifically, our method achieves the highest AUC of 6.0% on MIT-States, surpassing CANet by 0.6%. On the UT-Zappos, we achieve the highest HM of 49.3%, outperforming CANet by 2.0%. Although ProLT has a slight disadvantage on the C-GQA dataset, it remains competitive with the state-of-the-art methods, achieving an HM of 14.4%. As for the CLIP-based approaches, ProLT has produced remarkable outcomes. Unlike DFSP, our method avoids the incorporation of extra self-attention or cross-attention mechanisms. Despite this, we excel across all three datasets, attaining an HM of 38.2% on MIT-States and 49.4% on UT-Zappos. These results underscore the compatibility of ProLT when combined with CLIP.

Ablation Study

In this section, we verify that each of these modules plays an active role by ablating each of its parts on UT-Zappos with ResNet-18. The results are shown in Tab. 3 and Tab. 4.

Method	$\eta = 0$	$p = 0$	AUC	HM	S.	U.	Sta.	Obj.
GCN.	✓	×	29.4	44.2	59.9	63.0	43.4	70.0
	×	✓	31.3	47.3	60.8	60.5	43.7	73.6
	×	×	32.3	48.4	61.7	62.4	45.9	73.2
FC.	✓	×	28.9	44.1	60.1	62.9	44.1	73.2
	×	✓	32.7	48.3	61.8	64.6	45.8	74.1

Table 4: Ablation results for each component on UT-Zappos. $p = 0$ deontes we remove the prior in Eq. 14, ✓ indicates setting p or η to 0, and × indicates the opposite.

Attribute Prior versus Class Prior: As mentioned above, we use the attribute prior in place of the class prior due to the attribute imbalance. To further validate this, we replaced Eq. 11 and Eq. 14 using class prior, shown in Tab. 3. To make the results more robust, we tested two different prototype learners, *i.e.*, the GCN from the CGE (**?**) and the FC layers. The results in Tab. 3 indicate that incorporating a class prior yields improvements over the baseline. We attribute this enhancement mainly to **?**, the class sizes of datasets are not solely identical. However, ProLT exhibits a substantial advantage over the other methods, which demonstrates the more dominant influence of potential attribute imbalances in CZSL.

Effect of Components: We eliminate the effects of each component by adjusting the hyper-parameters η in Eq. 11 and the attribute prior in Eq. 16 to verify the role played by each component. In Tab. 4, we set η to 0 to convert Eq. 11 to a vanilla cross entropy loss and the inference phase is converted to same as CGE. For $p = 0$, we remove the attribute prior in inference phase. We also tested on both prototype learners. We can observe that each part of the ablation leads

Figure 3: Qualitative results on MIT-States (first row), UT-Zappos (second row) and C-GQA (third row), where the left part contains the top-3 results contains correct predicts, and the rights contains the top-3 predicts do not contain correct predicts. The label is indicated in black above the image, with correctly predicted results indicated in blue and incorrect ones in red.

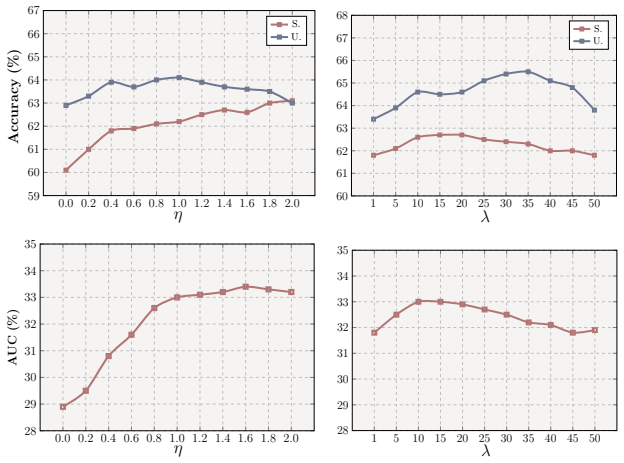

Figure 4: Influence of hyper-parameters on UT-Zappos about best seen (S.), best unseen (U.) and AUC.

to a decrease in outcome, with $\eta = 0$ being the most significant. This reflects the effectiveness of our method.

Hyper-Parameter Analysis

Our method primarily comprises the subsequent hyper-parameters: 1) logit-adjusting factor (η), and 2) factor about the sample distribution (λ). We test on the UT-Zappos under various hyper-parameters based on ResNet-18, shown in Fig. 4. For η, the best AUC are observed when $\eta = 1.6$, and the gap between seen and unseen begins to decrease as η increases. Concerning λ, the outcomes are documented within the interval $\lambda \in [1.0, 50.0]$ with increments about 5.0. The pinnacle value for the seen class is observed at 20.0, and 35.0 for unseen class. Overall, these hyper-parameter settings yield results characterized by minimal fluctuations, thus underscoring the robustness of our methodology.

Qualitative Results

Qualitative results for unseen compositions, accompanied by the top-3 predictions when we use ResNet-18 as backbone, are displayed in Fig. 3. Concerning MIT-States, we argue that certain erroneous predictions as partially justifiable. For instance, the phrase *tiny dog*, for which the model's incorrect predictions involve *small dog* and *tiny animal*, exhibits a high degree of semantic similarity. A similar phenomenon can be observed for the *brown chair* in C-GQA. For UT-Zappos, ProLT's limitation in fine-grained classification persists. An illustrative example is the outcomes for *leather boot.M*, our approach encounters challenges in making nuanced differentiations within the category of boots.

Conclusion

This paper presents from an experimental analysis aimed at revealing the concealed proximate long-tail distribution issue within CZSL. In our work, CZSL is transformed into an underlying proximate class imbalance problem, and the logit adjustment technique is employed to refine the posterior probability for individual classes. Diverging from conventional methods for handling long-tailed distributions, the introduced attribute prior is derived from the model's sample estimation of visual bias. Experimental results demonstrate that our approach attains state-of-the-art outcomes without necessitating the introduction of supplementary parameters.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) under the Grant No. 62371235.

Appendix

Supplementary Experiments and Details with Vision-Language Model

Implementation Details (Supplemental)

In this section, we provide details of the setup of ProLT when using CLIP to learn visual and semantic embeddings.

Visual Representations and Semantic: We employ the pretrained CLIP Vit-L/14 model as both our image and test encoder. Regarding semantics, we employ a learnable soft prompt *[v1][v2][v3][state][object]*, following the approach of **?**, where *[v1][v2][v3]* represent the learnable content. To embed attributes such as state or object, we compute the average embedding value for each composition containing the corresponding state or object.

Implementations and Hyper-Parameters: The three prototype learners, $\mathcal{P}_s, \mathcal{P}_o$, and $\mathcal{P}_y$, adhere to the configuration detailed in Sec. 4.3, except for the omission of GloVe (**?**). Similarly, the three visual embedders, $\mathcal{V}_s, \mathcal{V}_o$, and $\mathcal{V}_y$, remain consistent with the specifications in Sec. 4.3. We train the entire model using the Adam optimizer (**?**) on two NVIDIA GTX 3090 GPUs, while configuring the batch size as 16. The other hyper-parameter configurations remain consistent with those in main text.

Ablation Study with CLIP

Following Sec. 4.5, we conducted an identical experiment on CLIP to validate the effectiveness of ProLT. As demonstrated in Tab. A.5, we compare the outcomes on UT-Zappos (**?**) under three scenarios: without incorporating any priors but using a model-ensemble method, with the inclusion of class priors, and with the inclusion of attribute priors. Similarly, the results demonstrate the beneficial impact of incorporating the attribute prior. In comparison to the direct utilization of the class prior, our approach leads to a rise of 1.9% in AUC and 2.0% in HM.

Moreover, we conduct a comparative analysis by removing the attribute prior during both the training and testing phases. Referencing Tab. A.6, when $\eta = 0$, indicating our methods is changed to a simple common embedding space method like **?**, which led to a significant drop in results. A significant enhancement is observed when these are combined, similar to the findings in Tab. 4. Collectively, the aforementioned experiments substantiate the favorable impact of ProLT on CLIP.

Method	Prior	AUC	HM	S.	U.	Sta.	Obj.
CLIP	N.	33.6	46.5	63.4	69.2	50.9	79.5
	C.P.	34.2	47.4	64.1	68.1	51.2	77.8
	A.P.	36.1	49.4	66.0	70.1	52.6	79.4

Table A.5: A comparison of different priors for Eq. 11 and Eq. 14 when using CLIP as image and text encoder on UT-Zappos. **N.:** no prior is introduced, using pure ensemble method. **C.P.:** class prior from datasets is utilized. **A.P.:** attribute prior is incorporated.

Method	$\eta = 0$	$p = 0$	AUC	HM	S.	U.	Sta.	Obj.
CLIP	✓	×	31.7	45.9	62.7	66.2	48.8	75.7
	×	✓	35.1	47.8	65.0	69.3	52.1	79.9
	×	×	36.1	49.4	66.0	70.1	52.6	79.4

Table A.6: Ablation results for each component on UT-Zappos when using CLIP as image and text encoder. $p = 0$ deontes we remove the prior in Eq. 14, ✓ indicates setting p or η to 0, and × indicates the opposite.

Additional Experiments and Further Information

In this section we add some detailed information from Sec. 4 as well as perform some additional experiments. All experiments are performed with ResNet-18 (**?**) as the backbone.

Training Details

Early Stopping: As mentioned in Sec. 3.6, ProLT requires that $\mathcal{C}_s$ and $\mathcal{C}_o$ be trained together first using $\mathcal{L}_{ic}$. In this process we simply employ an early stopping strategy on the validation set. We trained these module for a maximum of 50 epochs and use AUC for early-stopping. After $\mathcal{C}_s$ and $\mathcal{C}_o$ training is complete, it starts outputting attribute priors and co-training with $\mathcal{C}_y$. This process we adopt the same early stopping strategy on the validation set. We set the maximum of 1000 epochs and also use AUC for early-stopping.

Word Embeddings	*AUC*	HM	S.	U.	Sta.	Obj.
GloVe	33.4	49.3	62.7	64.0	46.1	74.2
Word2Vec	32.4	48.8	63.1	64.9	45.6	75.0
Fasttext	32.5	48.4	63.0	62.6	45.4	74.9
GloVe+Word2Vec	33.0	49.4	63.1	63.8	45.4	74.1
Fasttext+Word2Vec	33.3	49.0	63.2	64.9	45.5	75.5

Table A.7: Results on UT-Zappos using different word embedding.

Hyper-Parameter Selection: Hyper-parameter selection involves grid-search on the validation set. For architectural parameters, we explore the 1) hidden layer count for $\mathcal{V}_s$, $\mathcal{V}_o$, and $\mathcal{V}_y$ within the range $0, 1, 2$, and 2) hidden layer count for $\mathcal{P}_s, \mathcal{P}_o$, and $\mathcal{P}_y$ within the same range. Concerning optimization, such as learning rate, we adopt the configuration from **?** without extensive modifications. For the remaining hyper-parameters, we search for $\eta \in [0.0, 2.0]$ with a step of 0.2, $\lambda \in [5, 50]$ with a step of 5 for UT-Zappos, and $\lambda \in [10, 200]$ with a step of 10 for MIT-States and C-GQA. Additionally, we perform a search for $\tau \in [0.02, 0.2]$ with an increment of 0.02, encompassing the value $\tau = \{0.005, 0.01\}$. For the choice of word embedding, we search the word embedding of 1) GloVe (**?**), 2) Word2Vec (**?**), 3) Fasttext (**?**), 4) GloVe+Word2Vec and (5) Fasttext+Word2Vec.

Dimension	*AUC*	HM	S.	U.	Sta.	Obj.
256	32.7	47.9	61.4	63.5	45.0	74.0
512	33.1	48.9	62.2	63.7	45.0	73.6
1024	33.4	49.3	62.7	64.0	46.1	74.2
2048	32.3	48.0	61.7	63.7	44.9	74.3
4096	32.2	48.3	61.9	64.4	46.3	74.7

Table A.8: Reults on UT-Zappos using different hidden layer settings.

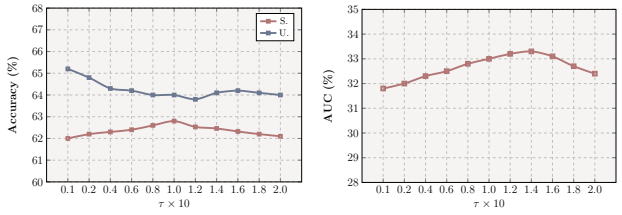

Figure A.5: Influence of τ on UT-Zappos about best seen (S.), best unseen (U.) and AUC.

Further Experiments of Hyper-Parameters

embeddings. Illustrated in Fig. A.5, we present the outcomes achieved across various τ values on UT-Zappos. The peak AUC emerges at $\tau = 0.14$ and the difference between seen and unseen is minimized at $\tau = 0.1$. Likewise, we present outcomes utilizing diverse word embeddings on UT-Zappos, detailed in Tab. A.7. ProLT excels when employing GloVe, yet generally, variations in word vectors exhibit minimal impact. Concerning the varying dimension configurations, we document the outcomes obtained using dimensions $256, 512, 1024, 2048, 4096$ for the hidden layers in three classifiers, as indicated in Tab. A.8. Notably, we discern that a hidden layer dimension of 1024 consistently yields optimal results. However, when employing dimensions of 2048 or 4096, we posit that the inferior performance could result from the propensity of higher-dimensional hidden layers to manifest overfitting on seen classes.

Explanation of Importance Sampling

to estimate the attribute prior for unseen classes. The specifics of this approach are outlined in this section. During this procedure, we introduce an auxiliary proposal distribution to aid in creating an approximate estimation, *i.e.*, the distribution of the seen attribute prior $k(s, o)$. Therefore, the estimation of the prior for unseen classes can be represented as follows:

$$\frac{1}{n}\sum_{i}^{n}\frac{p(\mathbf{x}_i)\hat{k}_{\mathbf{x}}(s,o)}{k(s,o)}. \tag{A.21}$$

replaced by λ, which is a hyper-parameter. This is owing to the unavailability of direct access to the data distribution for the test set. As the posterior can be obtained only for individual samples during testing, we set n to 1 in practice. This approach yields significant variance due to an inadequate sample size. Consequently, we posit that it should be integrated with the another information.

Method	$sp = 0$	$\lambda \to \infty$.	AUC	HM	S.	U.	Sta.	Obj.
GCN.	✓	✓	31.3	47.3	60.8	60.5	43.7	73.6
	✓	×	31.8	47.9	61.9	62.3	45.9	72.8
	×	✓	31.7	48.1	61.8	63.0	45.3	73.4
	×	×	32.3	48.4	61.7	62.4	45.9	73.2
FC.	✓	✓	32.7	48.3	61.8	64.6	45.8	74.1
	✓	×	32.8	48.5	62.0	64.0	46.0	74.3
	49.0	62.5	63.2	45.2	74.6			
×	×	33.4	49.3	62.7	64.0	46.1	74.2	

$= 0$ denotes we set the seen attribute prior in Eq. 16 to 0. And $\lambda \to \infty$ denotes we remove the probability from Importance Sampling. ✓ indicates true, and × indicates false.

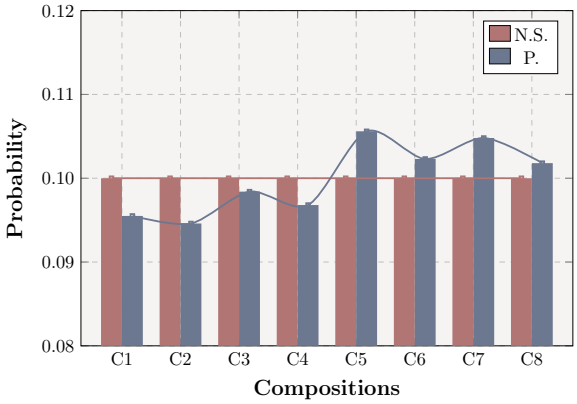

Figure A.6: An example of posterior and prior probabilities for various compositions using our method, where **N.S.** denotes the class prior, and the **P.** denotes the probabilities of $p(y|\mathbf{x})$. Our adjusted posterior provides a more balanced distribution compared to Fig. 1.

Further Ablation Study on Inference

approach on UT-Zappos that combining the seen attribute prior with Importance Sampling, as detailed in Sec. 3.5. Tab. **??** presents the outcomes where we nullify the probability estimated by Importance Sampling via setting $\lambda \to \infty$ in Eq. 16, and the results when we set the seen attribute prior to 0. It is worth noting that with seen attribute prior set to 0, we generalize the probability of Importance Sampling to the seen class for consistency. We conducted experiments on two embedding functions to ensure robustness following Tab. 3. From the table, we can observe that the introduction of the two respectively brings about an improvement in AUC, HM relative to the baseline, while it is not significant in the rest of the metrics. In addition, the combination of the two usually creates complementarities, suggesting that they are not mutually exclusive.

Why Our Method Works

more balanced distribution stemming from $\mathcal{C}_y$. Furthermore, it becomes apparent that $\mathcal{C}_y$ exhibits a preference for compositions characterized by significant visual bias in Fig. 1. But the incorporation of the prior, as defined in Eq. 14, can mitigates this distinction.

by harmonizing both a prior and a posterior during the inference phase.