Imaginary-Connected Embedding in Complex Space for Unseen Attribute-Object Discrimination

Chenyi Jiang, Shidong Wang, Yang Long, Zechao Li, Haofeng Zhang and Ling Shao, *Fellow, IEEE*

Abstract—Compositional Zero-Shot Learning (CZSL) aims to recognize novel compositions of seen primitives. Prior studies have attempted to either learn primitives individually (non-connected) or establish dependencies among them in the composition (fully-connected). In contrast, human comprehension of composition diverges from the aforementioned methods as humans possess the ability to make composition-aware adaptation for these primitives, instead of inferring them rigidly through the aforementioned methods. However, developing a comprehension of compositions akin to human cognition proves challenging within the confines of real space. This arises from the limitation of real-space-based methods, which often categorize attributes, objects, and compositions using three independent measures, without establishing a direct dynamic connection. To tackle this challenge, we expand the CZSL distance metric scheme to encompass complex spaces to unify the independent measures, and we establish an imaginary-connected embedding in complex space to model human understanding of attributes. To achieve this representation, we introduce an innovative visual bias-based attribute extraction module that selectively extracts attributes based on object prototypes. As a result, we are able to incorporate phase information in training and inference, serving as a metric for attribute-object dependencies while preserving the independent acquisition of primitives. We evaluate the effectiveness of our proposed approach on three benchmark datasets, illustrating its superiority compared to baseline methods. Our code is available at https://github.com/LanchJL/IMAX.

Index Terms—Compositional Zero-Shot Learning, Compositionality, Visual-Attribute, Complex Space, Open-World Classification

1 Introduction

Objects manifest in different attributes such as varying shapes, colors, and materials. Recognizing these attributes is a skill that people acquire [1], [2], [3] while machines find it challenging to master. The reason for this difficulty lies in the fact that supervised learning requires treating every composition of attributes and objects as a new class. The sheer magnitude of possible compositions makes it impractical to collect and label them all. Previous methods have proposed Compositional Zero-Shot Learning (CZSL) [4], [5] as a potential solution to this challenge. This method is capable of generalizing to unseen compositions by learning the seen attributes and objects.

Prior to commencing, we encourage the reader to contemplate the following: as a "model pre-trained on large amounts of data", how would a human envision the visual characteristics of an `black swan` based on images of `black cat` and `white swan`? Now, what if we modify the scenario to imagine an `old dog` created from an `old car` and an `cute dog`? In the case of the former, color is a feature that is constant across compositions, allowing humans to cover `swan` directly with `black` to perceive a `black swan`. Such primitives have a clear visual concept and show minor variations in different compositions. In contrast, the concept of `old` is highly abstract and manifests differently in `car` and `dog`. When considering the concept of `old` in various scenarios, human associations must integrate with the inherent characteristics of the other primitives within the composition because these primitives dynamically shift in visual representation depending on the composition.

Based on the characteristics described above, attributes can be broadly classified into consistent and dynamic types, and humans use two distinct modes of cognition to infer these attributes in unseen composition. This grants us to make dynamic choices based on the specific sample at hand, or we can refer to this mode of reasoning as composition-aware adaptation. Furthermore, the aforementioned two modes of human cognition have paved the way for the development of the two prevailing approaches to CZSL.

Illustrated in Fig. 1, one of the prevalent approaches aligns with the first of the above human cognition, which entails the learning of primitives (attributes and objects) within compositions [6], [7], [8]. For example, these methods are based on the decomposing of primitives `white` and `table`

Manuscript received 00-00, 2023; revised 00-00, 2023.
This work was supported in part by the National Natural Science Foundation of China under the Grant Nos. 62371235 and 62072246, in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant No. KYCX24_0681, and in part by the Key Research and Development Plan of Jiangsu Province (Industry Foresight and Key Core Technology Project) under Grant No. BE2023008-2. (Corresponding author: Haofeng Zhang.)
Chenyi Jiang, Zechao Li and Haofeng Zhang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China. Haofeng Zhang is also with the State Key Laboratory of Intelligent Manufacturing of Advanced Construction Machinery, Nanjing University of Science and Technology, Nanjing, 210094, China. E-mail:{jiangchenyi, zechao.li, zhanghf}@njust.edu.cn
Shidong Wang is with the School of Engineering, Newcastle University, Newcastle upon Tyne, NE17RU, United Kingdom. (e-mail: shidong.wang@newcastle.ac.uk)
Yang Long is with the Department of Computer Science, Durham University, Durham, DH13LE, United Kingdom. (e-mail: yang.long@ieee.org)
Ling Shao is with the UCAS-Terminus AI Lab, University of Chinese Academy of Sciences, Beijing, 100049, China. (e-mail: ling.shao@ieee.org)

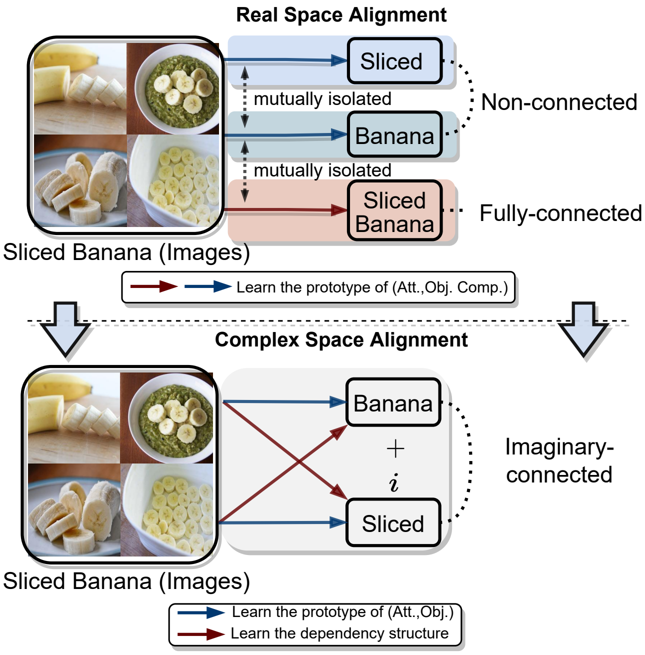


Fig. 1. Illustration of the motivation of the proposed method. Real Space Alignment: Two distinct real-space approaches (non-connected and fully-connected) that align with different modes of human cognition. The former learns primitives directly, while the latter learns compositionally dependent structures. Nonetheless, these approaches remain in three isolated measures, limiting their ability to make composition-aware adaptation as humans do. Complex Space Alignment: We unify the above three isolated measures into a single measure of the complex space employing imaginary-connected embeddings.

to distinguish a `white table`. Since attribute and object are embedded into two isolated measures, we refer to these methods as **non-connected** approaches. Another approach aligns with the second mode of human cognition [9], [10], allowing models to learn globally consistent compositional embeddings by leveraging the regularization of dependency structures between attributes and objects. In these methods, attributes are seamlessly integrated with objects within a unified framework, treating each composition as an independent class and disregarding the underlying shared concepts between seen and unseen classes. We classify these methods as **fully-connected** approaches.

However, as discussed above, humans possess the capability to composition-aware adaptation between these two modes of cognition depending on the context, a flexibility not present in the aforementioned methods. In CZSL, certain attributes (*e.g.*, `white, black, crushed`) are more compatible with the first approach, while others (*e.g.*, `huge, old, ancient`) align better with the second approach, or some samples necessitate a combination of both. While some prior research [11], [12] has attempted to merge the aforementioned methods for prediction, these isolated branches fail to interact and influence each other's inference. This paper aims to investigate an approach that integrates the aforementioned frameworks while closely aligning with the composition-aware adaptation of the human mind.

We drew inspiration from the representation of particle attributes in physics [13], [14]. These methods adopted the strategy of expressing the position and attribute of an object simultaneously with a complex number, and we can regard the compositions within CZSL as analogous concepts. We follow the above structure to establish an embedding structure where the attribute embedding remains **imaginary-connected** with the object embedding (`obj.` $+\, i$`att.`) in complex space, we refer to this as an imaginary-connected compositional embedding, shown in the bottom of Fig. 1. We treat each composition as an independent class rather than as a subset of the primitives within the composition. The attributes and objects belonging to the two independent measures are expressed as real and imaginary numbers. Embedding in complex space facilitates the knowledge transfer between two independent measures, attribute, and object, through phase information while maintaining their orthogonality. This unification enables the model to adaptively integrate non-connected and fully-connected inference strategies, thereby more closely mimicking human cognition.

The embedding mentioned above requires the decomposing of attributes and objects from visual features. Recent approaches [11], [15] commonly employ attention mechanisms to achieve this. Building upon this foundation, we leverage both spatial and channel locations to decompose the primitive features of a specific location, which we refer to as the **A**ttention-**G**uided **V**isual Decoupler (AGV). Nevertheless, as previously discussed, some attributes are highly dynamic and can extensively intertwine with visual features. Without sufficient data, models with fixed parameters struggle to learn attribute prototypes that bias visually with different compositions, which impairs their ability to generalize effectively across various compositions. To address this issue, we suggest employing an **O**bject-**G**uided **A**ttribute Extraction (OGA) module to facilitate attribute separation. Specifically, we integrate the standard deviation of the visual feature with respect to the object prototype to capture the visual bias influenced by the attribute. Subsequently, we combine this information with a specific local region to produce the decoupled attribute feature.

Apart from addressing seen classes, our proposed method facilitates generalization to unseen classes through the construction of an **A**ffinity-based **P**seudo **D**istribution (APD) in the complex space. Thus, the semantic information of the unseen compositions is incorporated into the training by leveraging the affinities among the compositions. Building upon the aforementioned configuration, we propose our framework, where embeddings are **IMA**ginary-connected in Comple**X** Space for Unseen Attribute-Object Discrimination (**IMAX**). This framework provides a more straightforward approach compared to prior methods as it avoids the need for constraints on input samples [8], [11] or the pre-construction of graph-based datasets [5]. This framework can be trained to utilize the semantics of both seen and unseen classes, showcasing its capability to conduct inference by dynamically integrating primitives and composition dependencies concurrently.

This paper introduces multiple significant contributions: 1) We present a novel method for attribute and object identification in complex space, enabling knowledge transfer between the independent measures of attribute and object through phase information, while preserving their orthogonality. 2) We model the impact of attributes by the bias between visual feature and object prototype and fuse with localized features to decompose attribute features that are deeply entangled with the objects. 3) Our proposed method is extensively evaluated through experiments and demonstrates superior performance compared to state-of-the-art methods on various challenging benchmark datasets.

2 Related Work

2.1 Compositionality

Compositionality is that objects or concepts can be understood through the combination of their constituent parts, essentially decomposing an observation into its primitives [16], [17]. Recent studies have explored compositionality in various tasks. At the visual level, several studies [18], [19], [20] have demonstrated improved generalization to diverse classes by learning prototypes of shared primitives between classes, these primitives are usually certain regions or features of the object. Applications at the semantic level have emerged, with recent studies [21], [22] employing Large Language Models to decompose class names into multi-perspective descriptive semantics, which serve as semantic primitives. These primitives can be accurately and intuitively captured by visual-semantic models, such as CLIP [23], thus directly providing interpretability for the model inference process. With objectives that differ slightly from previous approaches, CZSL extends the aforementioned concepts by imposing stricter constraints on compositional forms. CZSL employs objects and attributes as shared primitives across classes, emphasizing the ability to generalize to novel classes without exposure to training samples.

2.2 Zero-Shot Learning (ZSL)

The objective of ZSL is to transfer knowledge from seen classes to unseen classes [24], [25], [26], [27]. This generalization is primarily accomplished through semantic descriptions [28] or manually defined semantic attributes [29], [30]. From a compositional point of view, each class is decomposed into multiple primitives (hundreds or thousands), facilitating knowledge transfer across classes. For example, multiple attribute prototypes are utilized in APN [31] to extract the corresponding attributes from the localities of visual features. [32] leverages object detection [33] to decompose attribute-related regions in the feature to align with attributes. Unlike CZSL, ZSL methods seldom consider the visual bias introduced by the interaction of primitives as a major challenge. This is because these primitives generally correspond to more fixed semantic content, making ZSL often a regression task for attribute prediction. This disparity directly contributes to the difficulty of applying ZSL methods to CZSL.

2.3 Compositional Zero-Shot Learning (CZSL)

CZSL extends ZSL by incorporating compositionality, where each class is decomposed into two primitives: an object and an attribute. Unlike ZSL, the primitives in CZSL present greater visual complexity, because of the intricate visual interrelations among these primitives.

In response to this challenge, various studies have proposed different approaches, such as learning primitive classifiers and then combining them for the final inference [4], [34], [35], [36]. These approaches are referred to as non-connected methods. In this instance, causal inference has been introduced into CZSL for the decoupling of visual samples [6], [37], [38]. [7] decouples objects and attributes from images using Siamese networks, while [8] proposes a novel decoupling structure and uses the decoupled features to synthesize unseen class samples. These methods directly classify primitives in a composition, but neglecting the dependencies between them makes it difficult to handle the challenges of dynamic attributes.

The opposite approach considers the dependency structure between attributes and objects, treating each composition as a separate class, we refer to them as fully-connected methods. [10], [39] concatenate the word vectors of attributes and objects and embed them in a joint embedding space for classification. [5], [9] introduce Graph Convolutional Networks in CZSL, which use graph embeddings for compositions. Recent research treats attributes as direct conditions influencing object features. For example, [12] uses objects as conditions for generating attributes, while [15] emphasizes attribute-specific visual features by incorporating hierarchical guidelines within the visual attributes. These strategies aim to prompt the model to learn dependencies between primitives. However, unlike the non-connected approaches, they cannot focus on the shared primitives between classes.

With the recent advance in pre-trained vision-language models like CLIP [23], many methods show more prominent competitiveness. For instance, [40] introduces soft prompts [41] to mitigate the visual bias of primitives. [42] integrate language features with image features to reduce the domain gap between seen and unseen sets. [43] utilizes language-informed distribution to capture the intra-class diversity and inter-class correlation between primitives. These methods aim to fully leverage the zero-shot generalization capabilities of CLIP, benefiting from its large-scale pre-training data, and extend to more complex CZSL scenarios. As a method not tied to a specific visual or textual encoder, IMAX can similarly leverage the visual-semantic connections in CLIP for alignment in the complex space.

Many of the methods described above demonstrate different forms of modeling of compositions. For example, non-connected approaches [6], [7], [38] treat composition as a category mapping of primitives and do not perform direct inference on them. While the fully-connected approach differs, [44] view an attribute as a conditional operator, transforming it into a transformation operation linked to an object. In contrast, [5] conceptualizes their relationship using graphs. Additionally, [12] models the changes in attributes across different targets through conditional generation. The primary distinction between IMAX and these methods is its consideration of attributes as imaginary conditions associated with objects. This approach facilitates the learning of independent primitive prototypes and their inherent connections in a single framework. Consequently, IMAX integrates the advantages of both methods into a unified prediction system.

2.4 Complex Space

Various tasks in deep learning incorporate the use of complex space. For instance, [45] defines each relationship

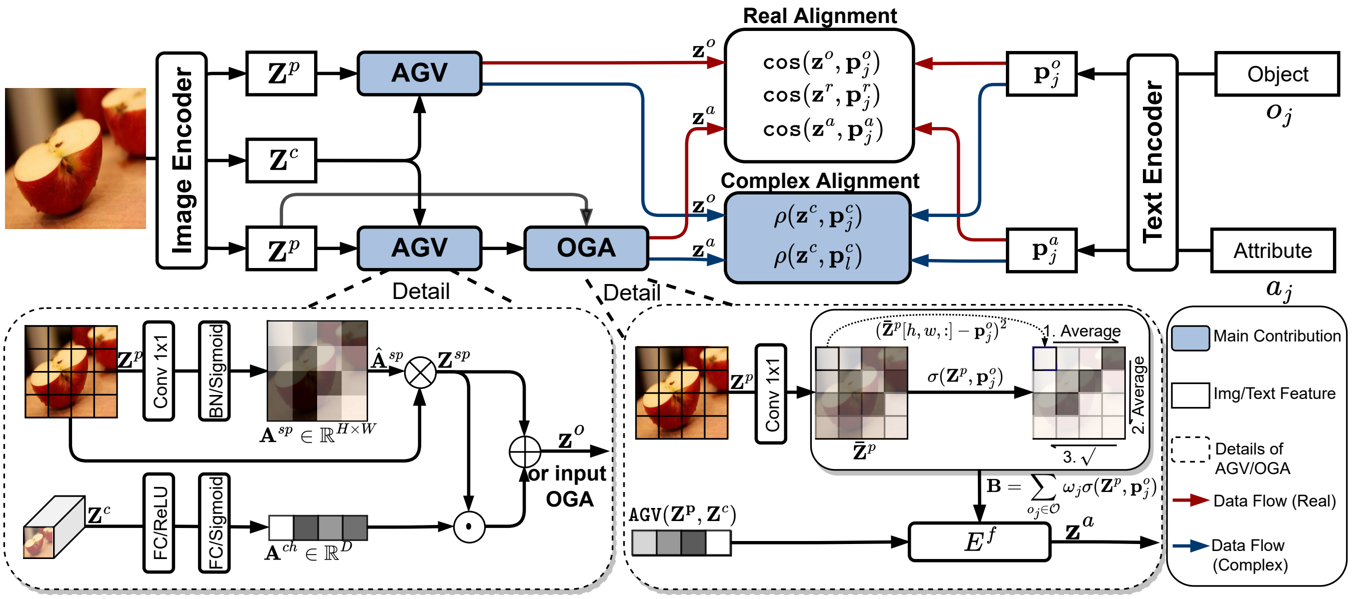


Fig. 2. The data flow of our approach is illustrated. The input visual features $\mathbf{Z}^p$ and $\mathbf{Z}^c$ are decomposed by the AGV and OGA modules to derive $\mathbf{z}^a$ and $\mathbf{z}^o$. The class labels a_j and o_j are encoded as $\mathbf{p}^a_i$ and $\mathbf{p}^o_j$. Both features are then projected into complex space and real space for visual-semantic alignment. Specific details of the AGV and OGA are also presented. AGV takes $\mathbf{Z}^p$ and $\mathbf{Z}^c$ as inputs and generates local activation maps for spatial and channel dimensions, denoted as $\mathbf{A}^{sp}$ and $\mathbf{A}^{ch}$, respectively, where matrix multiplication is represented by $\otimes$, element-wise product by $\odot$, and element-wise summation by $\oplus$. As for the OGA module, $\mathbf{Z}^p$ undergoes a $\mathtt{conv}_{1\times 1}$ operation with noise filtering to obtain $\bar{\mathbf{Z}}^p$. The variance is then computed at each spatial location using the object prototype $\mathbf{p}^o_j$ as the visual bias and weighted through Eq. 7. $\mathbf{B}$ is fused with the output of the AGV module to obtain the final attribute feature. The visual-semantic alignment for complex space and real space is demonstrated in Fig. 3.

within the knowledge graph as a rotation from a source entity to a target entity in a complex vector space. In the context of Domain Generalization [46], several methods represent style using imaginary numbers. For instance, [47] considers the domain as a style in the frequency domain and introduces filters to mitigate its impact. Additionally, [48] leverages the frequency domain to enhance data in the target domain. Similarly, [49] examine the quantitative phase variation of normalization through mathematical derivation of the Fourier transform formula and eliminate style while preserving content through spectral decomposition. Although these methods aim to model style using the Fourier transform in the frequency domain, they do not directly involve classification in complex space. To the best of our knowledge, no existing method represents embeddings in complex space for the CZSL task.

3 Approach

In CZSL, every image in the dataset is annotated with an attribute and an object label, such as `old` (attribute) and `dog` (object), respectively. The task of CZSL requires recognizing novel compositions consisting of attributes and objects from the seen compositions. The complex interdependence between attributes and objects adds to the challenge of this task. Below we provide a full description of our approach.

3.1 Task Formulation

In CZSL, labels are often compositions of attributes and objects. We denote the possible attributes and objects as a and o respectively and use $y = (a, o)$ to represent the corresponding labels of images. The label set can be expressed as $\mathcal{C} = \mathcal{A} \times \mathcal{O} = \{(a, o)|a \in \mathcal{A}, o \in \mathcal{O}\}$. An input sample $\mathbf{x} \in \mathcal{X}$ is an image in the space $\mathcal{X}$. The set of training images can be denoted as $\mathcal{X}_t \subset \mathcal{X}$ and the corresponding labels as $\mathcal{C}_t \subset \mathcal{C}$. Therefore, $\mathcal{T} = \{(\mathbf{x}, y)|\mathbf{x} \in \mathcal{X}_t, y \in \mathcal{C}_t\}$ is the training set used to train a mapping function $\mathcal{X} \to \mathcal{C}_n$, where $\mathcal{C}_n \subseteq \mathcal{C}$ represents the test compositions. It is important to note that $\mathcal{C}_n$ may contain compositions that are not present in $\mathcal{C}_t$. Depending on the relationship between $\mathcal{C}_n$ and $\mathcal{C}$, the following tasks can be defined: 1) Generalized CZSL following [50], which can also be denoted by Closed-World CZSL (CW-CZSL), where $\mathcal{C}_n \subset \mathcal{C}$ and the test class includes both feasible existing seen and unseen classes. Most recent works [5], [7], [8] have followed this setting. (2) Open World CZSL (OW-CZSL) following [10], which requires $\mathcal{C}_n \equiv \mathcal{C}$. It is a more challenging task due to the presence of a large number of interference compositions. Notably, most recent works [5], [10] on CZSL consider the set of unseen compositions in $\mathcal{C}_n$ is assumed to be known a priori. Consequently, IMAX retains this setup.

This paper addresses both CW-CZSL and OW-CZSL settings. Particularly, OW-CZSL poses additional challenges as the number of compositions tested ($\mathcal{C}_n$) exceeds the number of compositions in training ($\mathcal{C}_t$) by a significant margin. Moreover, OW-CZSL involves a high proportion of infeasible compositions (over 90% in MIT-States [51]). Consequently, the model needs to go beyond recognizing attributes and objects alone and also consider the feasibility between them.

3.2 Method Overview

For an input sample $\mathbf{x}$ and its label is $y_j = (a_j, o_j)$, we utilize a pre-trained backbone, such as ViT-B-16 (ViT-B),

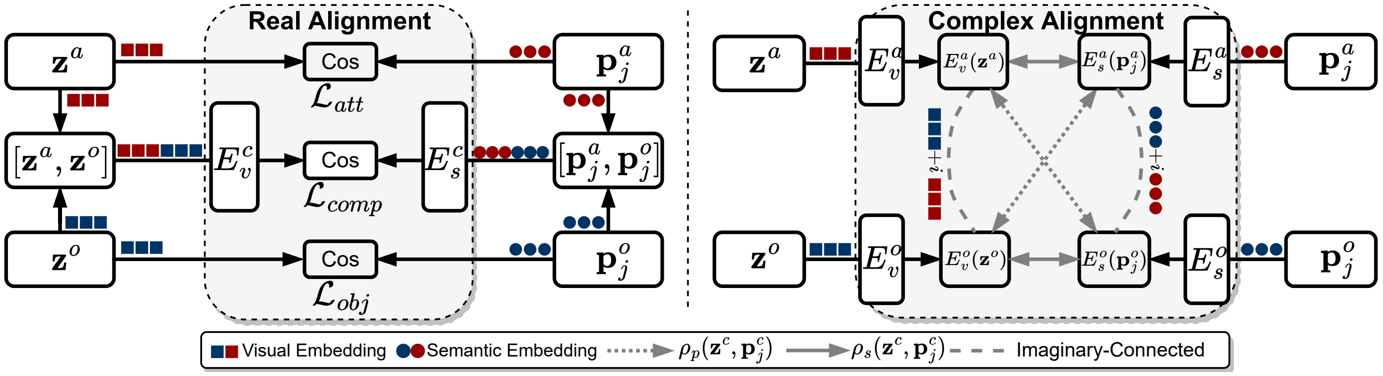


Fig. 3. Illustration of visual-semantic alignment in real space and complex space. In real alignment, the visual and semantic embeddings for objects and attributes are projected into their respective real spaces to compute cosine similarity. Additionally, we combine $\mathbf{z}^o$ and $\mathbf{z}^a$ into compositional forms that align with the compositional semantics after embedding, serving as an auxiliary loss. In complex alignment, $\mathbf{z}^a$ and $\mathbf{z}^o$ are embedded in the complex space and represented as the imaginary-connected form. The semantic embedding is also expressed in the same form in the complex space. The computation of visual-semantic similarity in complex space incorporates two elements: the direct similarity between the primitives ($\rho_s(\mathbf{z}^c, \mathbf{p}_j^c)$) and the utilization of phase information to model the attribute-object dependency ($\rho_p(\mathbf{z}^c, \mathbf{p}_j^c)$).

ResNet-18 (R-18) [52] or ViT-L-14 (ViT-L) [53] in CLIP [23], to extract visual features, as depicted in Fig. 2. If ViT is used as the backbone, we can obtain a patch token, and reshape it to $\mathbf{Z}^p \in \mathbb{R}^{H\times W\times D}$. We can also obtain a class token (*i.e.*, `[CLS]`) $\mathbf{Z}^c \in \mathbb{R}^D$, where D representing the channel number.

When R-18 is used as the backbone, we extract the visual feature map after layer 4, represented by $\mathbf{Z}^p \in \mathbb{R}^{H\times W\times D}$. To align with ViT's methodology, we consider the features obtained after applying `Avgpool` as a vector similar to the class token, represented as $\mathbf{Z}^c = \texttt{Avgpool}(\mathbf{Z}^p) \in \mathbb{R}^D$. It should be noted that the values of H, W, and D differ when using R-18 and ViT, this setting is solely intended to facilitate the subsequent presentation.

For the semantic branch, we utilize GloVe [54] followed by two FC-layers or CLIP text encoder [23] to embed the label $y_j = (a_j, o_j)$ to latent space, thus we can obtain the word embeddings which work as the attribute and object prototypes, *i.e.*, $\mathbf{p}_j^a, \mathbf{p}_j^o \in \mathbb{R}^D$.

The aforementioned aligns principally with the previous CZSL approach [11], [15], serving as a rudimentary module. In subsequent modules, previous approaches typically use attention mechanisms [8], [11], [15] or linear functions [6], [7] to decompose or map visual features. In IMAX, because of the deep entanglement of attributes and objects, we propose the OGA (Sec. 3.4) module by decomposing features about attributes through the standard deviation of visual features and object prototypes, outside the use of attention mechanisms (AGV, Sec. 3.3). In the process of learning compositional prototypes, we overcome the limitations of the real space, it lacks the dynamics to combine dependencies and primitives for classification. Instead, we leverage a complex space to the synergy between the above two, *i.e.*, in the form of imaginary-connected compositional embedding in complex space (Sec. 3.6). Pseudo-distributions for unseen classes can be constructed in complex spaces by leveraging the affinities between compositions, enabling guidance for generalization (Sec. 3.8).

3.3 Attention-Guided Visual Decoupler (AGV)

Compared to attributes, objects have explicit texture features as well as boundaries in the image, which is easier to perceive directly in the local regions. AGV encompasses attention mechanisms based on convolution kernels [55], targeting the precise activation of pertinent spatial regions. Unlike previous methods, AGV incorporates a channel activation module [56] that originates from the feature's global information. The intersection of these two components yields our ultimate decoupling outcome.

Specifically, when provided with a visual feature $\mathbf{Z}^p$, we train filters to evaluate its degree of association with the corresponding primitives in the spatial location of the feature. This evaluation is then fed into a batch normalization (`BN`) layer and activated using a `Sigmoid` function to highlight the relevant regions. As a result, we obtain the spatial activation map $\mathbf{A}^{sp}$:

$$\mathbf{A}^{sp} = \texttt{Sigmoid}(\texttt{BN}(\texttt{Conv}_{1\times1}(\mathbf{Z}^p))) \in \mathbb{R}^{H\times W}, \quad (1)$$

the $\texttt{conv}_{1\times1}$ serves as the primitive-aware filters and its output dimension is 1. Then we use this activation map to activate the original visual features, *i.e.*,

$$\mathbf{Z}^{sp} = \hat{\mathbf{A}}^{sp} \otimes (\hat{\mathbf{Z}}^p)^\top \in \mathbb{R}^D, \quad (2)$$

where $\hat{\cdot}$ denotes flattened features, *i.e.*, $\hat{\mathbf{A}}^{sp} \in \mathbb{R}^{HW}$, $\hat{\mathbf{Z}} \in \mathbb{R}^{D\times HW}$. $\otimes$ denotes the matrix multiplication and $\top$ denotes transpose. Due to the deep entanglement of attributes with objects, we like to perform a weighting of global visual features over channels. Thus, we similarly generated channel-based activation maps via class token:

$$\mathbf{A}^{ch} = \texttt{Sigmoid}(\texttt{SE-Block}(\mathbf{Z}^c)), \quad (3)$$

where `SE-Block` is derived from [56], consisting of an FC-layer for dimensionality reduction, a ReLU function, and another FC-layer for dimensionality enhancement along the channel axis. In the end, we combine the channel activation maps obtained above with spatial activation features and introduce a residual connection:

$$\mathbf{z} = \texttt{AGV}(\mathbf{Z}^p, \mathbf{Z}^c) = \mathbf{Z}^{sp} \odot \mathbf{A}^{ch} + \mathbf{Z}^{sp} \in \mathbb{R}^D, \quad (4)$$

where $\odot$ denotes the element-wise product. $\mathbf{z}$ is the output feature of this module. We use this module to decouple object features, *i.e.*, $\mathbf{z}^o = \texttt{AGV}_o(\mathbf{Z}^p, \mathbf{Z}^c)$, where the $\texttt{AGV}_o$ denotes the AGV model is for object. As for attributes, a further process is required to achieve composition-aware adaptation.

3.4 Object-Guided Attribute Extraction (OGA)

The embeddings $\mathbf{p}_j^o$ of objects in visual space are viewed as object prototypes, representing the most fundamental and essential concept of objects across the samples. Taking inspiration from methods for extracting styles in domain generalization [46], [57]. To represent the bias generated by the combination of attribute and object, we compute the standard deviation of the spatial feature $\mathbf{Z}^p$ along with its corresponding object prototype $\mathbf{p}_j^o$, and it is this bias that is the most significant visual variation of the attribute. In this way, we aim to accomplish the extraction of attributes from the visual bias that occurs in the object.

Since the object to which the samples correspond is not known a priori, we assign weights to these standard deviations based on the similarity of the features to all object prototypes, which can be calculated as:

$$w_j = \texttt{Softmax}(\cos(\mathbf{z}^o, \mathbf{p}_j^o)). \tag{5}$$

The standard deviation between the visual feature $\mathbf{Z}^p$ and the object prototype $\mathbf{p}_j^o$ is utilized as the visual bias introduced by the composition:

$$\sigma(\mathbf{Z}^p, \mathbf{p}_j^o) = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(\bar{\mathbf{Z}}^p[h,w,:] - \mathbf{p}_j^o)^2}, \tag{6}$$

where $\bar{\mathbf{Z}}^p \in \mathbb{R}^{H\times W\times D}$ is obtained by pass the $\mathbf{Z}^p$ through a $\texttt{conv}_{1\times1}$ in order to filter out background noise, $[h,w,:]$ denotes the index $[h,w,:]$ in feature $\bar{\mathbf{Z}}^p$. Next, we apply a weight to all standard deviations using the weight w_j:

$$\mathbf{B} = \sum_{o_j\in\mathcal{O}} w_j\sigma(\mathbf{Z}^p, \mathbf{p}_j^o) \in \mathbb{R}^D. \tag{7}$$

Additionally, we incorporate local attention information as a complementary aspect, and subsequently combine it with the weighted standard deviation to generate our ultimate decoupled attribute features:

$$\mathbf{z}^a = E^f([\texttt{AGV}_a(\mathbf{Z}^p, \mathbf{Z}^c), \mathbf{B}]) \in \mathbb{R}^D, \tag{8}$$

where $[\texttt{AGV}_a(\mathbf{Z}^p, \mathbf{Z}^c), \mathbf{B}]$ denotes the concatenation of the two vectors, E^f is a function consisting of two FC-layers for fusing the vectors, and $\texttt{AGV}_a$ is an AGV model for attributes.

3.5 Training the Real Space Decoupling Modules

Based on the above modules, we complete the decomposing of visual features in real space. As shown on the left of Fig. 3, we use a vanilla cross-entropy loss function to guide the training process of these modules:

$$\mathcal{L}_{att} = -log\frac{exp\{\cos(\mathbf{z}^a, \mathbf{p}_j^a)/\tau\}}{\sum_{y_{j'}\in\mathcal{C}_t} exp\{\cos(\mathbf{z}^a, \mathbf{p}_{j'}^a))/\tau\}}, \tag{9}$$

$$\mathcal{L}_{obj} = -log\frac{exp\{\cos(\mathbf{z}^o, \mathbf{p}_j^o)/\tau\}}{\sum_{y_{j'}\in\mathcal{C}_t} exp\{\cos(\mathbf{z}^o, \mathbf{p}_{j'}^o))/\tau\}}, \tag{10}$$

where τ is temperature coefficient, $y_j = (a_j, o_j)$ is the label of the input sample. And $\texttt{cos}$ denotes cosine similarity. Following previous work [15], we avoid loss of information by reconstructing the composition in the process, which can be formulated as:

$$\mathbf{z}^r = E_v^c([\mathbf{z}^a, \mathbf{z}^o]),\ \mathbf{p}_j^r = E_s^c([\mathbf{p}_j^a, \mathbf{p}_j^o])), \tag{11}$$

where E_v^c and E_s^c are functions that include two FC-layers which are also trained using a vanilla cross-entropy loss:

$$\mathcal{L}_{comp} = -log\frac{exp\{\cos(\mathbf{z}^r, \mathbf{p}_j^r)/\tau\}}{\sum_{y_{j'}\in\mathcal{C}_t} exp\{\cos(\mathbf{z}^r, \mathbf{p}_{j'}^r))/\tau\}}. \tag{12}$$

The objective function for real space can be summarized as follows:

$$\mathcal{L}_{real} = \mathcal{L}_{att} + \mathcal{L}_{obj} + \mathcal{L}_{comp}. \tag{13}$$

3.6 Imaginary-Connected Embedding

As discussed in Sec. 1, it is plausible to directly utilize the aforementioned modules for classifying the samples in the real space. However, in the real-space approach, the attributes and objects exist as completely isolated measures, inability to dynamically combine dependencies between primitives. Therefore, we propose employing imaginary-connected compositional embeddings in the complex space to represent the compositions. In complex spaces, attributes are considered as imaginary conditions that are attached to the objects, rather than having a direct connection with them. This approach allows the unifying of several independent measures in real space into a common measure in complex space, which can capture the inherent interdependencies of composition, while also preserving shared information between both seen and unseen classes.

Specifically, we embed the $\mathbf{z}^a$, $\mathbf{z}^o$ into complex space, *i.e.*,

$$\mathbf{z}^c = E_v^o(\mathbf{z}^o) + iE_v^a(\mathbf{z}^a), \tag{14}$$

where E_v^o and E_v^a are FC-layers, and i denotes imaginary unit. Specifically, we convert the $\mathbf{z}^a$ into imaginary numbers, while maintaining the $\mathbf{z}^o$ as real numbers. Subsequently, these two features are amalgamated to form a complex vector, *i.e.*, imaginary-connected embeddings.

From this, we perform a similar operation for the semantic branch:

$$\mathbf{p}_j^c = E_s^o(\mathbf{p}_j^o) + iE_s^a(\mathbf{p}_j^a), \tag{15}$$

E_s^o and E_s^o are also embedding networks same to E_v^o and E_v^a. The above two complex vectors compute the similarity $\rho(\mathbf{z}^c, \mathbf{p}_j^c)$ in complex space, *i.e.*,

$$\rho(\mathbf{z}^c, \mathbf{p}_j^c) = \frac{\langle\mathbf{z}^c, \mathbf{p}_j^c\rangle}{\langle\mathbf{z}^c, \mathbf{z}^c\rangle\,\langle\mathbf{p}_j^c, \mathbf{p}_j^c\rangle}, \tag{16}$$

where $\langle\mathbf{z}^c, \mathbf{p}^c\rangle$ denotes the Hermitian inner product between $\mathbf{z}^c$ and $\mathbf{p}^c$. As a result, this calculation contains a real number for the similarity score and an imaginary number for the complex phase factor, *i.e.*,

$$\rho(\mathbf{z}^c, \mathbf{p}^c) = \rho_s(\mathbf{z}^c, \mathbf{p}_j^c) + i\rho_p(\mathbf{z}^c, \mathbf{p}_j^c). \tag{17}$$

For simplicity, we discuss below in the case of $|\mathbf{z}^c| = 1$ and $|\mathbf{p}_j^c| = 1$. As a similarity score,

$$\rho_s(\mathbf{z}^c, \mathbf{p}_j^c) = E_o^v(\mathbf{z}^o)^\top E_o^s(\mathbf{p}_j^o) + E_a^v(\mathbf{z}^a)^\top E_a^s(\mathbf{p}_j^a). \tag{18}$$

which is used to directly measure the visual-semantic match for objects and attributes, plays a role similar to the cosine similarity of the real space in the previous methods. As for the complex phase factor, it is obtained by:

$$\rho_p(\mathbf{z}^c, \mathbf{p}_j^c) = E_a^v(\mathbf{z}^a)^\top E_o^s(\mathbf{p}_j^o) - E_o^v(\mathbf{z}^o)^\top E_a^s(\mathbf{p}_j^a), \qquad (19)$$

mathematically, it is used to provide the complex phase factor necessary for achieving complete similarity in Eq. 18. Fundamentally, a smaller $\rho_p(\mathbf{z}^c, \mathbf{p}_j^c)$ indicates a smaller complex angle between the two vectors, suggesting a higher level of global coherence. In IMAX, we expect a larger $\rho_s(\mathbf{z}^c, \mathbf{p}_j^c)$ between visual features and semantics, and a smaller $\rho_p(\mathbf{z}^c, \mathbf{p}_j^c)$. In complex space, IMAX's visual-semantic alignment of seen classes is optimized using Eq. 20, after which we provide a detailed discussion and explanation.

$$\mathcal{L}_s = -log\frac{exp\{(\rho_s(\mathbf{z}^c, \mathbf{p}_j^c) - \rho_p(\mathbf{z}^c, \mathbf{p}_j^c))/\tau\}}{\sum_{y_{j'} \in \mathcal{C}_t} exp\{(\rho_s(\mathbf{z}^c, \mathbf{p}_{j'}^c) - \rho_p(\mathbf{z}^c, \mathbf{p}_{j'}^c)/\tau\}}. \qquad (20)$$

3.7 Discussion on Complex Space Alignment

As discussed in Sec. 1, we group the attributes into two types: consistent attributes and dynamic attributes. Consistent attributes (*e.g.*, color, textures) represent attributes that maintain constant visual features regardless of composition, making them easier to disentangle directly from visual features. In contrast, dynamic attributes (*e.g.*, `old`, `ripe`, `huge`) vary based on the object, which means $\mathbf{z}^a$ may associated with the object. Although the OGA and AGV modules effectively disentangle features, dynamic attributes may exist directly with the object itself, and it's hard to distinguish using lightweight models.

In CW-CZSL, an ideal consistent attribute pertaining to completely orthogonal to the object, *i.e.*, $\rho_p(\mathbf{z}^c, \mathbf{p}_j^c) \to 0$. On the other hand, $\mathbf{z}^a$ from dynamic attributes exhibit some degree of similarity to the $\mathbf{p}_j^o$. We aim to utilize this similarity to guide the learning of $E_a^v(\mathbf{z}^a)$ related to these attributes, encouraging them to deviate from their primitive semantics and emphasize the relationship with the object. To achieve this, we expect to maintain close values between the $E_a^v(\mathbf{z}^a)^\top E_o^s(\mathbf{p}_j^o)$ and $E_o^v(\mathbf{z}^o)^\top E_a^s(\mathbf{p}_j^a)$. Both scenarios indicate our expectation of a lower $\rho_p(\mathbf{z}^c, \mathbf{p}_j^c)$. Thus, in the phase information, we accomplish the establishment of the attribute-object dependency in this simple way.

During the inference stage of OW-CZSL, infeasible compositions often lack complete orthogonalization between attributes and objects, as well as the corresponding attribute-object dependency. Consequently, this leads to a larger value for the phase factor. Therefore, we argue that phase information can aid in assessing the feasibility of such compositions.

As a result, our optimal state for the training phase is for the visual-semantic sample pairs to have the smaller complex phase factor while achieving higher similarity. From this, we can introduce the objective function at this stage of the process:

3.8 Affinity-based Pseudo Distribution (APD)

While the above work explores composition further, another key to the CZSL task lies in the generalization of unseen classes, which is hard to achieve by relying on the extremely small amount of data available. Many approaches have been proposed in ZSL to guide the direction of training for unseen classes, mainly by minimizing entropy [58], [59]. In CZSL, we begin by calibrating the distribution of word embeddings for unseen classes using the semantic similarity between the compositions of seen and unseen classes. To constrain the pseudo-distribution of unseen classes, we introduce inter-class affinity for re-weighting. We have considered inter-class affinity, which refers to the composition's overlapping semantic relations (same objects or attributes between different compositions).

Following the previous process, word vectors for unseen classes are also embedded in complex space, which is $\mathbf{p}_l^a$ and $\mathbf{p}_l^o$, we denote the complex form of it by $\mathbf{p}_l^c$, l denotes $y_l = (a_l, o_l)$, assuming we have a total of m_s seen classes and m_u unseen classes, we have $l = m_s+1, m_s+2, ..., m_s+m_u$. For complex vectors $\mathbf{p}_j^c$ of seen class y_j, we employ the similarity as the pseudo distribution on unseen classes, $\rho_s(\mathbf{p}_j^c, \mathbf{p}_l^c)$ denotes the similarity of the j-th seen class to the l-th unseen class.

Thereafter, each sample is assigned an unseen class pseudo-label based on affinity. First we need to define $y_{k_j} = (a_{k_j}, o_{k_j})$, where,

$$k_j = \text{argmax}_{m_s+1 \le k' \le m_s+m_u} \ \rho_s(\mathbf{p}_j^c, \mathbf{p}_{k'}^c), \qquad (21)$$

y_{k_j} denotes the unseen word embedding that is most affinity to $y_j = (s_j, o_j)$. Using y_{k_j} as a label for unseen classes enables the construction of a pseudo-distribution for correction purposes. Nevertheless, in the CZSL scenario, we examine the existence of overlapping semantics between classes and wish to model the inter-class relationships of unseen classes using this concept. Here, we propose a method named Affinity-based Re-weighting, which is defined as follows:

$$\mu_{j,l} = \begin{cases} 1-\epsilon & y_l = y_{k_j} \\ 0 & s_l, o_l \notin y_{k_j} \\ \pi(\epsilon) & otherwise \end{cases}, \qquad (22)$$

where $\epsilon \in (0,1)$ is a hyper-parameter that is used to adjust the weights, to avoid interference caused by infeasible compositions, we usually set a larger ϵ in OW-CZSL. $\pi(\epsilon)$ is to adjust the weights of the remaining affinity compositions,

$$\pi(\epsilon) = \frac{\epsilon}{K-1}, \qquad (23)$$

where K is the number of unseen compositions with affinity to y_{k_j}, *i.e.*, affinity here means between two compositions, there exist the same primitive (attributes or objects).

In summary, we use this pseudo-label to warm up the model's ability to generalize to unseen classes, this is accomplished by a loss function similar to Eq. 20:

$$\mathcal{L}_u = -\sum_{l=m_s+1}^{m_u} \mu_{j,l} log\frac{exp\{(q_s(\mathbf{z}^c, \mathbf{p}_l^c) - q_p(\mathbf{z}^c, \mathbf{p}_l^c))/\tau\}}{\sum_{y_{l'} \in \mathcal{U}} exp\{(q_s(\mathbf{z}^c, \mathbf{p}_{l'}^c) - q_p(\mathbf{z}^c, \mathbf{p}_{l'}^c))/\tau\}}. \qquad (24)$$

Ultimately, the modules in the complex space are trained by the following functions:

$$\mathcal{L}_{complex} = \mathcal{L}_s + \alpha\mathcal{L}_u, \qquad (25)$$

where α is a weighting coefficient that balances the two losses.

TABLE 1
Dataset statistics for CZSL, *a*: number of attributes, *o*: number of objects, *sp*: seen compositions, *up*: unseen compositions, *i*: images.

			Training		Validation			Test		
Dataset	a	o	sp	i	sp	up	i	sp	up	i
MIT-States [51]	115	245	1262	30k	300	300	19k	400	400	13k
UT-Zappos [51]	16	12	83	23k	15	15	3k	18	18	3k
C-GQA [5]	413	674	5592	27k	1252	1040	7k	888	923	5k

3.9 Inference

At the inference phase, we feed the input into AGV and OGA modules to get $\mathbf{z}^a$, $\mathbf{z}^o$ and $\mathbf{z}^r$, and the test labels are also embedded in the corresponding space, for class y_j we have $\mathbf{p}^a_j, \mathbf{p}^o_j$ and $\mathbf{p}^r_j$. We first compute its similarity to the semantics in real space following [11]: $s_r(\mathbf{x}, y_j) = \cos(\mathbf{z}^a, \mathbf{p}^a_j) \cdot \cos(\mathbf{z}^o, \mathbf{p}^o_j) + \cos(\mathbf{z}^r, \mathbf{p}^r_j)$.

In complex space, we embed the above vectors to get $\mathbf{z}^c$ and $\mathbf{p}^c_j$. And we continue to introduce complex phase information into the classification phase, which is $s_c(\mathbf{x}, y_j) = \rho_s(\mathbf{z}^c, \mathbf{p}^c_j) - \rho_p(\mathbf{z}^c, \mathbf{p}^c_j)$. We compute our prediction score by synthesizing these scores in different spaces:

$$\hat{y} = \mathrm{argmax}_{y_j \in \mathcal{C}_n} \beta s_r(\mathbf{x}, y_j) + s_c(\mathbf{x}, y_j), \tag{26}$$

where β is a hyper-parameter for synthesizing the scores.

4 Experiments

This section describes the experimental data, evaluates the criteria, and undertakes experiments to validate IMAX.

4.1 Datasets

In compliance with previous research on CZSL and the interest of a fair comparison, our method is evaluated on three main benchmark datasets, *i.e.*, MIT-States [51], UT-Zappos [60], and C-GQA [5], as shown in Tab. 1.

MIT-States [51]. Image labels in MIT-States are generated through automated labeling of image-text search, and [37] found that leads to a high level of noise in both training and test sets. Nonetheless, we contend that the MIT-States dataset remains suitable for evaluating model performance in extremely noisy environments and can be compared fairly to previous studies. The dataset includes around 53,000 images, comprising 245 object classes and 115 attribute classes. We employ the standard split of [62], in the closed-world setting, the dataset consisted of 1262 seen classes, and the validation and test set consisted of 300/400 unseen classes, respectively. For open-world scenarios, we follow [10] who considers all possible 28175 pairs and about 26114 compositions of them are not present in any splits of the dataset. However, the OW-CZSL needs to consider the possibilities of these compositions.

UT-Zappos [60]. We evaluate our method on UT-Zappos, comprising around 33,000 fine-grained footwear images categorized into 16 attribute classes and 12 object classes. For the CW-CZSL, the training set contains 83 seen compositions, and the validation and test sets consist of 15/18 unseen compositions, respectively, as described in [62]. For open-world scenarios, the splits of UT-Zappos also follow [10], which contains 192 possible compositions, and about 76 are not in any of the splits of the dataset.

C-GQA [5]. [5] proposes a new dataset C-GQA for CZSL, which has less noise and more compositions than MIT-States, and we use it in our experiments. Following [9], we split the dataset into over 9,000 compositions, consisting of 5,592 seen and 1,938 unseen compositions in the CW-CZSL setting. The vast search space of C-GQA makes it the most challenging dataset among the three. In OW-CZSL, C-GQA remains the most challenging dataset, containing a search space of 280K possible compositions [9].

4.2 Evaluation Metrics

We follow the evaluation metrics outlined in [62] for both closed-world and open-world settings. The main metrics evaluated are (1) Area Under the Curve (AUC); (2) Accuracy of Seen and Unseen compositions ($\mathcal{A}^S$, $\mathcal{A}^U$); (3) Harmonic Mean (HM) ($\mathcal{A}^H$), where $\mathcal{A}^H = (2 \times \mathcal{A}^S \times \mathcal{A}^U)/(\mathcal{A}^U + \mathcal{A}^S)$; Metric (1) evaluate the areas between accuracy on seen and unseen compositions with different bias terms [5], [8], [62]. Metrics (2) directly evaluate the accuracy when testing only seen or unseen classes. Metric (3) is a common metric for generalized ZSL following [50] and is used to comprehensively evaluate the performance of the model on seen and unseen classes.

4.3 Implementation Details

Models. IMAX is compatible with various image and text encoders. We primarily utilize pre-trained ResNet-18 [52] and ViT-B-16 [53] for image encoding, while GloVe [54] is employed for text encoding to maintain consistency with baseline methods [11], [15]. Additionally, IMAX can be implemented using the pre-trained CLIP ViT-L-14 model for both image and text encoding [23]. The pre-trained word embeddings of CLIP for `a photo of` are used to initialize three prefixes for input semantics. When implemented with CLIP, the composition semantics, such as `a photo of a [mashed] [banana]`, are encoded as $[\mathbf{p}^a_j, \mathbf{p}^o_j]$ for Eq. 11, rather than using a direct concatenation of the two encoded vectors from attribute and object. In all implementations, the embedding layers in IMAX are two-layer MLPs connected by a ReLU activation function.

Training setup. IMAX is trained on two NVIDIA RTX A6000 GPUs using the ADAM optimizer [64], with a learning rate of 5×10^{-5} and a batch size of 256 (64 when fine-tuning the visual backbone and 16 for CLIP), employing an early stopping strategy. For MIT-States, α and β are set to 1×10^{-5} and 0.5, respectively; for UT-Zappos, these values are 5×10^{-6} and 0.7. For C-GQA, α is set to 1×10^{-4}, and β is set to 100. In CW-CZSL, ϵ is consistently set to 0.25 across all three datasets, while in OW-CZSL, the values are 0.4, 0.65, and 0.45. Additionally, the temperature coefficient τ is set to 0.1, 0.125, and 0.01 for the respective datasets.

4.4 Results

We evaluate IMAX in both CW-CZSL and OW-CZSL settings. We report IMAX results using different implementations: R-18 [52], ViT-B [53], and CLIP ViT-L [23] models. For experiments using R-18 as the backbone, we compare outcomes with recent baselines such as OADis [8] and CANet [12]. When using ViT-B as the backbone, we select

TABLE 2
The results of IMAX and baselines on MIT-States, UT-Zappos, and C-GQA in CW-CZSL. We compare our method with others on AUC, best harmonic mean, seen and unseen accuracy ($\mathcal{A}^H$, $\mathcal{A}^S$, and $\mathcal{A}^U$). The best AUC and $\mathcal{A}^H$ are shown in bold.

Method	Backbone	MIT-States				UT-Zappos				C-GQA			
		AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	*AUC*	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	*AUC*	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$
With frozen R-18													
LE+ [4]	R-18	2.0	10.7	15.0	20.1	25.7	41.0	53.0	61.9	0.8	6.1	18.1	5.6
AttOp [44]	R-18	1.6	9.9	14.3	17.4	25.9	40.8	59.8	54.2	0.7	5.9	17.0	5.6
TMN [61]	R-18	2.9	13.0	20.2	20.1	29.3	45.0	58.7	60.0	1.1	7.5	23.1	6.5
SymNet [6]	R-18	3.0	16.1	24.4	25.2	23.9	39.2	53.3	57.9	2.1	11.0	26.8	10.3
CompCos [10]	R-18	4.5	16.4	25.3	24.6	28.7	43.1	59.8	62.5	2.6	12.4	28.1	11.2
CGE [5]	R-18	5.1	17.2	28.7	25.3	26.4	41.2	56.8	63.6	2.3	11.4	28.1	10.1
SCEN [7]	R-18	5.3	18.4	29.9	25.2	32.0	47.8	63.5	63.1	2.9	12.4	28.9	12.1
OADis [8]	R-18	5.9	18.9	31.1	25.6	30.0	44.4	59.5	65.5	3.0	13.3	30.2	12.5
CANet [12]	R-18	5.4	17.9	29.0	26.2	33.1	47.3	61.0	66.3	3.3	14.5	30.0	13.2
CoT [15]	R-18	**6.2**	19.6	30.8	26.8	31.5	46.4	62.1	64.2	**4.5**	**16.6**	33.1	16.6
IMAX*	R-18	**6.2**	**19.7**	31.0	26.5	**33.9**	**49.5**	64.3	64.8	**4.5**	16.4	32.7	16.9
With frozen ViT-B													
CompCos [10]	ViT-B	7.5	22.0	33.3	30.0	31.8	48.1	58.8	63.8	2.9	12.8	30.7	12.2
CGE [5]	ViT-B	7.3	21.3	33.5	28.6	34.5	48.5	61.6	70.0	3.8	15.0	32.3	14.9
OADis [8]	ViT-B	7.5	21.9	34.2	29.3	32.6	46.9	60.7	68.8	3.8	14.7	33.4	14.3
CANet [12]	ViT-B	**8.3**	22.3	36.3	30.7	36.3	49.5	64.0	72.0	4.9	17.5	34.4	17.3
ADE [11]	ViT-B	7.7	22.8	31.0	32.0	35.1	51.1	63.0	64.3	5.2	18.0	35.0	17.7
CoT [15]	ViT-B	7.8	23.2	34.8	31.5	33.7	48.5	60.2	65.0	5.1	17.5	34.0	18.8
IMAX*	ViT-B	8.0	**23.6**	35.2	30.1	**41.4**	**55.3**	68.0	72.3	**5.4**	**18.9**	35.4	18.6
With fine-tuning ViT-B													
CompCos [10]	ViT-B	7.8	22.4	36.3	30.4	39.0	53.3	65.6	67.8	4.8	16.7	38.4	16.6
CGE [5]	ViT-B	9.7	24.8	39.7	31.6	39.2	53.8	66.8	67.9	5.4	18.5	38.0	17.1
OADis [8]	ViT-B	10.1	25.2	39.2	32.1	33.0	48.7	62.4	68.7	7.0	20.1	38.3	19.8
CANet [12]	ViT-B	8.8	23.1	37.5	31.1	38.7	52.2	67.2	69.5	5.6	18.9	38.0	17.1
ADE [11]	ViT-B	6.7	20.1	33.5	28.1	38.1	53.6	65.0	66.7	5.2	18.7	34.2	17.9
CoT [15]	ViT-B	10.5	25.8	39.5	33.0	34.2	49.8	63.5	63.4	7.4	22.1	39.2	22.7
IMAX*	ViT-B	**10.9**	**26.1**	39.8	34.5	**42.0**	**57.4**	69.8	70.1	**7.6**	**23.4**	39.6	24.1
With frozen CLIP													
CLIP [23]	ViT-L	11.0	26.1	30.2	46.0	5.0	15.6	15.8	49.1	1.4	8.6	7.5	25.0
CoOp [41]	ViT-L	13.5	29.8	34.4	47.6	18.8	34.6	52.1	49.3	4.4	17.1	20.5	26.8
CSP [40]	ViT-L	19.4	36.3	46.6	49.9	33.0	46.6	64.2	66.2	6.2	20.5	28.8	26.8
DFSP (i2t) [42]	ViT-L	20.7	37.2	47.4	52.4	32.1	45.1	64.2	66.4	8.7	24.3	35.6	29.3
DFSP (BiF) [42]	ViT-L	20.8	37.7	47.1	52.8	33.5	47.1	63.3	69.2	9.0	26.2	36.5	32.0
DFSP (t2i) [42]	ViT-L	20.6	37.3	46.9	52.0	36.0	47.2	66.7	71.7	10.5	27.1	38.2	32.0
PLID [43]	ViT-L	**22.1**	39.0	49.7	52.4	38.7	52.4	67.3	68.8	11.0	27.9	38.8	33.0
IMAX*	ViT-L	21.9	**39.1**	48.7	53.8	**40.6**	**54.2**	69.3	70.7	**12.8**	**29.8**	39.7	35.8

CANet, CoT [15], and ADE [11] as main baselines and we conducted separate tests with both frozen and fine-tuned visual backbones. In addition, we compare several CLIP-based baselines [40], [42], [43] to demonstrate IMAX's capability to integrate with large-scale vision-language models.

Closed-world setting. As shown in Tab. 2, we compare IMAX with recently proposed baselines. On both visual backbones, IMAX outperforms previous SoTA methods on most datasets.

Although IMAX is constrained by the absence of class tokens when using the frozen R-18 backbone, this limitation does not significantly impact its performance. Compared to existing methods, IMAX improves the $\mathcal{A}^H$ by $+0.1\%$ and $+1.7\%$ on MIT-States and UT-Zappos, respectively. Our results on C-GQA closely align with the current SoTA, with a minor difference of (-0.2% on $\mathcal{A}^H$) and equality in AUC compared to CoT. Considering CoT utilizing multiple visual features at different scales, our method achieves close performance without the need for additional visual features. An intriguing observation arises from the comparison of IMAX with CANet, particularly its attribute generation conditioned on objects. We posit that this concept is inherently embedded in IMAX, wherein attributes are treated as conditions associated with objects. In comparison, we possess an advantage wherein phase information exhibits stronger generalization capabilities than conditional generation, given the absence of the introduction of additional parameters.

When employing ViT-B as the backbone, It can be observed that fine-tuning the visual backbone generally outperforms using frozen backbones. This suggests that,

TABLE 3
The results of IMAX and other baselines on MIT-States, UT-Zappos, and C-GQA in OW-CZSL.

Method	Backbone	MIT-States				UT-Zappos				C-GQA			
		AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	*AUC*	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	*AUC*	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$
With frozen R-18													
LE+ [4]	R-18	0.3	2.7	14.2	2.5	16.3	30.5	60.4	36.5	0.08	1.0	19.2	0.7
AttOp [44]	R-18	0.7	4.7	16.6	5.7	13.7	29.4	50.9	34.2	-	-	-	-
TMN [61]	R-18	0.1	1.2	12.6	0.9	8.4	21.7	55.9	18.1	-	-	-	-
SymNet [6]	R-18	0.8	5.8	21.4	7.0	18.5	34.5	53.3	44.6	0.43	3.3	26.7	2.2
CompCos^{cw} [10]	R-18	0.9	5.9	25.3	5.5	20.8	36.3	59.8	45.6	0.20	1.6	28.0	1.0
CompCos [10]	R-18	1.6	8.9	25.4	10.0	21.3	36.9	59.3	46.8	0.39	2.8	28.4	1.8
CGE [5]	R-18	0.7	4.9	29.6	4.0	21.5	38.0	58.8	46.5	0.30	2.2	28.3	1.3
Co-CGE [9]	R-18	2.0	10.1	26.4	10.4	21.3	38.1	60.1	44.3	0.37	2.6	28.7	1.6
KG-SP [39]	R-18	1.0	6.7	23.4	7.0	22.9	39.1	58.0	47.2	0.44	3.4	26.6	2.1
DRANet [63]	R-18	1.1	6.9	27.1	6.6	23.5	39.7	60.7	46.1	0.71	5.0	28.2	3.1
CANet [12]	R-18	1.5	8.1	32.0	7.3	22.2	37.7	58.8	50.3	0.47	3.2	30.0	2.0
IMAX*	R-18	**2.2**	**10.5**	29.7	10.8	**25.5**	**43.1**	61.3	52.5	**0.73**	**5.3**	28.8	2.6
With frozen ViT-B													
CompCos [10]	ViT-B	2.8	11.6	32.7	12.7	20.7	36.0	58.1	46.0	0.72	4.3	32.8	2.8
CGE [5]	ViT-B	1.0	6.0	28.0	5.9	23.5	40.0	60.6	47.0	0.81	4.8	32.7	3.2
OADis [8]	ViT-B	2.9	12.0	33.0	12.5	25.3	41.6	58.7	53.9	0.71	4.2	33.0	2.6
CANet [12]	ViT-B	2.6	10.7	38.2	9.7	27.5	43.2	63.1	54.7	0.91	5.0	34.2	3.1
ADE [11]	ViT-B	2.6	11.7	31.3	12.2	27.1	44.8	62.4	50.7	1.42	7.6	35.1	4.8
CoT [15]	ViT-B	3.1	12.4	36.7	11.8	27.9	44.0	63.3	53.3	1.22	6.5	35.0	4.3
IMAX*	ViT-B	**3.5**	**14.2**	35.1	13.5	**30.9**	**46.1**	68.1	60.8	**1.53**	**8.0**	35.3	4.9
With fine-tuning ViT-B													
CompCos [10]	ViT-B	3.1	12.8	31.0	14.2	25.1	41.3	62.5	49.8	1.14	6.0	34.0	4.1
CGE [5]	ViT-B	1.2	6.3	29.6	6.0	24.3	41.0	57.3	49.5	0.95	5.4	33.4	3.4
OADis [8]	ViT-B	3.0	12.8	33.6	12.3	26.7	43.4	60.0	52.9	1.22	6.2	36.2	4.0
CANet [12]	ViT-B	2.5	11.1	34.1	11.2	27.9	43.9	66.2	54.2	1.16	5.9	37.1	3.8
ADE [11]	ViT-B	2.6	11.9	31.3	12.3	28.1	45.4	64.3	49.8	1.48	7.9	35.9	4.7
CoT [15]	ViT-B	**4.2**	14.3	41.1	14.1	28.0	44.1	62.2	48.7	1.33	7.0	35.3	4.4
IMAX*	ViT-B	**4.2**	**14.4**	39.2	15.1	**32.4**	**47.3**	68.8	59.3	**1.59**	**8.3**	35.9	5.1
With frozen CLIP													
CLIP [23]	ViT-L	3.0	12.8	30.1	14.3	2.2	11.2	15.7	20.6	0.27	4.0	7.5	4.6
CoOp [41]	ViT-L	2.8	12.3	34.6	9.3	13.2	28.9	52.1	31.5	0.70	5.5	21.0	4.6
CSP [40]	ViT-L	5.7	17.4	46.3	15.7	22.7	38.9	64.1	44.1	1.20	6.9	28.7	5.2
DFSP(i2t) [42]	ViT-L	6.7	19.1	47.2	18.2	26.4	41.2	64.3	53.8	1.95	9.0	35.6	6.5
DFSP(BiF) [42]	ViT-L	6.7	19.2	47.1	18.1	27.6	42.7	63.5	57.2	2.39	10.6	36.4	7.6
DFSP(t2i) [42]	ViT-L	6.8	19.3	47.5	18.5	30.3	44.0	66.8	60.0	2.40	10.4	38.3	7.2
PLID [43]	ViT-L	7.3	20.4	49.1	18.7	30.8	46.6	67.6	55.5	2.50	10.6	39.1	7.5
IMAX*	ViT-L	**7.6**	**21.4**	**50.2**	**18.6**	**32.3**	**47.5**	68.4	57.3	**2.58**	**11.2**	38.7	7.9

TABLE 4
Comparison of the complexity between the IMAX method and baseline methods on C-GQA, evaluated based on the number of model parameters, FLOPS, and average runtime per sample. All methods utilize ViT-B as the backbone, with results excluding ViT-B's complexity indicated in red. Experiments are conducted on an NVIDIA RTX A6000 GPU.

Methods	Parameters ($\times 10^6$)	FLOPS ($\times 10^9$)	Run Time (ms)
CANet [12]	89.710 (4.063)	21.229 (4.366)	156.311 (152.196)
COT [15]	90.351 (4.704)	26.679 (9.816)	21.458 (17.221)
ADE [11]	109.840 (24.193)	22.454 (5.591)	114.684 (110.672)
IMAX*	97.469 (11.822)	29.638 (12.775)	25.960 (22.002)

despite the risk of overfitting within seen classes due to limited data, the bias in the backbone's pre-training data relative to the test scenarios remains the more dominant factor. As a result, IMAX improves the $\mathcal{A}^H$ by $+0.4\%$, $+4.2\%$ and $+0.9\%$ on the three datasets with a frozen ViT-B when compared with SoTA, respectively. When we fine-tune the backbone, IMAX also improves the $\mathcal{A}^H$ by $+0.3\%$, $+3.6\%$ and $+1.3\%$. While replacing the R-18 backbone with ViT-B improves most baselines' results, IMAX shows clear superiority across all three datasets. We attribute this to the adaptability of AGV to the ViT-B architecture, enhanced by available class tokens, which are crucial in fine-grained scenarios where decoupling visual features is essential for constructing complex visual-semantic relations.

CLIP-based experimental results demonstrate that IMAX can effectively adapt to different image and text encoders. IMAX represents a significant improvement on $\mathcal{A}^H$ com-

TABLE 5
Ablation study results of IMAX about the training and inference space. *CW* denotes the CW-CZSL setting, and *OW* denotes the OW-CZSL setting.

	Training	Inference	MIT-States				UT-Zappos				C-GQA			
			AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$
CW	Real	Real	6.2	19.6	32.1	27.5	33.8	48.2	64.8	64.5	4.7	16.4	33.4	17.9
	Real	Complex	6.6	20.5	33.9	27.6	37.1	52.8	64.6	69.7	4.7	17.5	35.5	16.3
	Complex	Real	7.0	21.1	34.6	29.6	35.3	50.8	62.8	67.5	4.3	16.2	33.2	16.5
	Complex	Complex	8.0	23.6	35.2	30.1	41.4	55.3	68.0	72.3	5.4	18.9	35.4	18.6
OW	Real	Real	1.7	9.6	33.9	9.1	23.9	40.0	64.9	44.6	0.83	5.1	31.5	3.1
	Real	Complex	2.1	12.5	34.7	13.8	29.8	45.7	67.9	60.3	1.17	6.9	34.7	4.3
	Complex	Real	1.8	10.3	34.2	10.4	25.1	42.2	63.6	59.2	0.96	5.7	35.6	3.0
	Complex	Complex	3.5	14.2	35.1	13.5	30.9	46.1	68.1	60.8	1.53	8.0	35.3	4.9

pared to vanilla CLIP [23], achieving $+13.0\%$, $+38.6\%$, and $+21.2\%$ higher performance across various benchmarks. Although CLIP itself exhibits strong zero-shot inference ability, its understanding of primitives in composition is still limited. Compared to the rest of the baselines designed for CZSL, IMAX outperforms the existing methods by $+0.1\%$, $+1.8\%$ and $+1.9\%$ in terms of $\mathcal{A}^H$. Compared to the multiple cross-attention and self-attention mechanisms in DFSP, IMAX achieves more prominent results with a more lightweight model structure. We attribute this to IMAX's deeper dissection of primitive relationships within the composition.

Open-world setting. Results for the OW-CZSL setting are presented in Tab. 3. IMAX continues to achieve leading performance on this setup, we consider the phase information in complex space to offer an enhanced measure of the attribute-object dependence, which suppresses the probability that unfeasible compositions be predicted in the inference phase. This proves beneficial for distinguishing between unfeasible compositions in this scenario.

In the context of utilizing R-18 as the backbone, we observe the improvement of $+0.4\%$ in terms of $\mathcal{A}^H$ metrics, respectively, compared to Co-CGE [9] on MIT-States. For UT-Zappos, our method aligns with the CW-CZSL setting, achieving 25.5% and 43.1% on AUC and $\mathcal{A}^H$, a performance nearly on par with SoTA ViT-Based methods and surpassing earlier approaches (CompCos [10], CGE [5]) on CW-CZSL. Notably, in the challenging C-GQA, rich in unfeasible compositions, our method leverages phase information in complex space, achieving a new SoTA with $+0.3\%$ improvement in $\mathcal{A}^H$ and $+0.02\%$ on AUC compared to DRANet [63]. Given the dataset's extensive compositions, this margin is deemed highly significant.

When implemented with ViT-B, similar trends can be observed as in CW-CZSL. IMAX improves $\mathcal{A}^H$ by $+1.8\%$, $+1.3\%$, and $+0.4\%$ on the three datasets with a frozen ViT-B, respectively. When fine-tuning the backbone, IMAX also improves $\mathcal{A}^H$ by $+0.1\%$, $+1.9\%$, and $+0.4\%$. It is noteworthy that ViT-B-based IMAX's results on UT-Zappos are still the most prominent; however, they are not as significant as those on CW-CZSL. This difference is attributed to the fact that the number of infeasible compositions in the OW-CZSL setting of UT-Zappos is much smaller compared to other datasets. As a result, inference on this content is less critical in this context, allowing some CW-CZSL-based methods to also achieve excellent performance on this dataset.

The results of implementing CLIP demonstrate that IMAX effectively integrates with CLIP to enhance reasoning in the OW-CZSL setting, addressing a limitation inherent to the CLIP model. The suppression of infeasible compositions through phase information allows IMAX to fully leverage the strong visual-semantic connections inherent in CLIP. Consequently, IMAX improves $\mathcal{A}^H$ by $+1.0\%$, $+0.9\%$, and $+0.6\%$, and AUC by $+0.3\%$, $+1.5\%$, and $+0.08\%$, respectively.

Combining the above results, we contend that, in open-world scenarios, complex space assumes a more crucial role compared to alternative methods. Earlier methods typically categorized compositions by solely decoupling visual features, a process that frequently falls short in determining the feasibility of a composition. Co-CGE assesses each composition by modeling a feasibility score and continually updating it, but this process is computationally intensive.

Computational efficiency. We compare the complexity of the models by analyzing the number of parameters and the computational overhead of each model. Here we focus on comparisons with the following baselines: COT [15], ADE [11] and CANet [12]. As shown in Tab. 4, we report the number of parameters and Floating Point Operations Per Second (FLOPS) for IMAX with these baseline methods. All compared methods use the same backbone architecture (ViT-B) and perform inference on C-GQA. In addition to reporting the parameter count, FLOPS, and runtime for the full model, we also present these results excluding ViT-B by replacing the feature extraction step with randomly generated vectors of the same size.

We observe that the number of parameters introduced by IMAX is the second lowest among the four methods, but it has higher FLOPS due to the similarity calculation in complex space during inference. However, IMAX does not suffer from significant runtime drawbacks, as we enhance time efficiency during the inference phase by utilizing matrix multiplication instead of extensive index lookups. The specific implementation details are available in our code. Overall, we consider that IMAX improves performance on the three benchmark datasets while maintaining model complexity within a reasonable range.

TABLE 6
Validation of AGV and OGA modules is performed on MIT-States, UT-Zappos, and C-GQA. ✓ indicates the module is included in IMAX, whereas × indicates it is not included.

	Attribute ($\mathbf{z}^a$)		Object ($\mathbf{z}^o$)	MIT-States				UT-Zappos				C-GQA			
	AGV	OGA	AGV	AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$
CW	×	×	×	6.2	20.1	31.2	28.7	34.1	50.2	61.7	66.3	4.9	17.8	34.0	16.8
	×	×	✓	6.3	20.5	31.4	29.1	35.2	51.3	63.7	68.3	4.9	18.0	35.1	16.5
	×	✓	×	7.4	22.0	33.6	29.7	37.1	52.9	65.3	68.5	5.1	18.4	35.3	18.0
	×	✓	✓	7.5	22.3	35.9	30.2	38.9	53.9	66.9	69.3	5.2	18.7	36.0	18.1
	✓	×	×	7.0	20.2	36.2	28.1	37.2	52.4	66.3	70.9	5.4	18.3	35.3	18.6
	✓	×	✓	7.2	20.9	37.2	28.9	38.6	53.4	67.5	71.5	5.3	18.4	35.6	19.1
	✓	✓	×	7.5	23.0	35.8	31.5	39.5	54.6	67.0	71.3	5.4	18.7	35.0	19.4
	✓	✓	✓	8.0	23.6	35.2	30.1	41.4	55.3	68.0	72.3	5.4	18.9	35.4	18.6
OW	×	×	×	2.6	13.0	32.4	12.9	24.7	42.8	62.6	54.5	1.28	6.9	33.4	4.1
	×	×	✓	2.9	13.2	33.7	13.3	25.7	43.3	63.4	56.8	1.32	7.0	33.9	4.0
	×	✓	×	3.1	13.7	34.3	13.6	27.3	44.2	65.7	58.6	1.40	7.4	34.6	4.5
	×	✓	✓	3.3	14.0	35.1	13.3	28.3	45.2	68.2	60.2	1.49	7.8	34.8	5.1
	✓	×	×	3.0	13.4	34.9	13.1	27.0	43.7	67.2	59.1	1.30	7.1	34.2	4.5
	✓	×	✓	3.1	13.5	35.4	12.9	27.6	44.5	67.5	59.9	1.36	7.3	34.7	4.3
	✓	✓	×	3.3	13.9	34.6	13.7	29.3	45.3	67.9	60.3	1.47	7.9	35.1	5.0
	✓	✓	✓	3.5	14.2	35.1	13.5	30.9	46.1	68.1	60.8	1.53	8.0	35.3	4.9

4.5 Ablation Study

This section presents an analysis of the validity of each module in the methodology. All experiments use frozen ViT-B as the backbone and are conducted on MIT-States, UT-Zappos, and C-GQA. We present experimental results for both CW-CZSL and OW-CZSL to assess the impact of each module in these distinct settings.

Complex space *vs* real space. As the most significant contribution of our work, we first verify the difference between visual-semantic alignment in complex space and real space. Specifically, we eliminate the imaginary terms derived from similarity computation during both training and inference, in this way we transform the entire method of inference and training into real space.

As shown in Tab. 5, we report results for removing complex spaces in the training phase as well as in the inference phase, respectively. The results demonstrate that visual-semantic alignment in complex space yields positive gains during both the inference and training phases. It is noteworthy that introducing complex phase information only in the inference phase can also result in a significant improvement in the OW-CZSL setting, about $+0.4\%, +5.7\%$, and $+0.34\%$ improvement on AUC across the three datasets, respectively. We speculate that this improvement stems from its capability to reduce the likelihood of retrieving less feasible compositions by discerning the interdependence of attributes and objects through the inherent generalization capability of the model. Overall, a more straightforward observation is that introducing complex spaces during both the training and testing phases yields better results, as it helps maintain consistency between these phases.

Effective of AGV and OGA. We also validate the roles of the AGV and OGA modules. To obtain the decomposed representations $\mathbf{z}^a$ and $\mathbf{z}^o$, an AGV module is employed for each branch, while the OGA module is additionally utilized for the attributes. We achieve the ablation of these two modules by replacing them with a simple linear structure, *i.e.*, `Avgpool` and a two-layer MLP, the results are shown in Tab. 6.

In both datasets, we observe a decrease in results when both modules are removed simultaneously. This suggests that having well-decomposed visual features remains a necessary condition for IMAX. Although the AGV module itself has positive effects, we find that OGA plays a more significant role. In CW-CZSL, the use of OGA alone yields results higher than using AGV alone. This confirms our hypothesis that relying solely on fixed parameters hinders the adaptation of biased visual representations of the same primitives from various compositions. On the other hand, OGA can use the object as a conditional prior, dynamically separating its attributes. Furthermore, we observe a significant improvement when both modules are introduced simultaneously for decoupling attributes. This suggests that the fusion output of AGV and OGA has a positive effect on attribute extraction. Overall, both AGV and OGA modules provide significant benefits across all three datasets, particularly on fine-grained datasets like UT-Zappos. These parameters facilitate a clearer delineation of categorization boundaries between primitives in the compositions.

Different components of AGV. We validate the performance of the AGV by disassembling each component ($\mathbf{A}^{sp}, \mathbf{A}^{ch}$) in the AGV module. These two components realize the decomposing of visual features in terms of spatial location as well as channel location, respectively. We report the experimental results for the CW-CZSL as well as OW-CZSL settings in Tab. 7.

The introduction of $\mathbf{A}^{ch}$ demonstrates noticeable improvements in both tasks. In CW-CZSL and OW-CZSL, compressing the channels on all three datasets gives a minor but consistent boost, *e.g.*, $+0.4\%$ and $+1.6\%$ on AUC for MIT-States and UT-Zappos in CW-CZSL. This supports our assertion that there is a substantial overlap between attributes and objects in terms of spatial location. Regarding $\mathbf{A}^{sp}$, although the enhancement from this module in OW-CZSL is not as pronounced as in CW-CZSL, it still acts as

TABLE 7
Ablation study results for the components in AGV. ✓ indicates that AGV includes this component, while × indicates the absence of this component in AGV.

	Spatial ($\mathbf{A}^{sp}$)	Channel ($\mathbf{A}^{ch}$)	MIT-States				UT-Zappos				C-GQA			
			AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$
CW	×	×	7.4	22.0	33.6	29.7	37.1	52.9	65.3	68.5	5.3	18.4	35.3	18.0
	✓	×	7.6	22.2	34.8	30.1	39.2	54.8	68.2	71.2	5.6	18.7	35.9	19.4
	×	✓	7.8	22.5	33.8	29.0	38.7	53.9	67.5	71.5	5.4	18.4	35.9	18.8
	✓	✓	8.0	23.6	35.2	30.1	41.4	55.3	68.0	72.3	5.4	18.9	35.4	18.6
OW	×	×	3.1	13.7	34.3	13.6	27.3	44.2	65.7	58.6	1.40	7.4	34.6	4.5
	✓	×	3.3	13.9	35.3	13.0	27.7	45.0	63.4	59.8	1.48	7.8	35.1	4.6
	×	✓	3.4	14.0	34.9	13.1	28.3	45.7	67.7	59.8	1.45	7.5	35.4	4.5
	✓	✓	3.5	14.2	35.1	13.5	30.9	46.1	68.1	60.8	1.53	8.0	35.3	4.9

TABLE 8
Effectiveness analysis of each component in APD, the experiment is conducted on MIT-States, UT-Zappos, and C-GQA.

	Baselines	MIT-States				UT-Zappos				C-GQA			
		AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$
CW	$\alpha \to 0$	7.4	21.8	36.7	27.5	37.1	53.0	68.6	67.5	5.0	17.3	34.1	17.1
	$\epsilon \to 1$	7.6	22.0	35.1	28.4	38.7	53.9	68.9	71.3	5.1	17.1	34.9	17.6
	$\epsilon \to 0$	7.7	22.6	35.3	29.3	41.2	54.8	67.0	72.0	5.2	18.3	35.0	18.5
OW	$\alpha \to 0$	3.1	13.8	35.9	11.4	27.9	44.2	68.3	57.3	1.36	7.4	36.4	2.5
	$\epsilon \to 1$	3.4	14.0	35.4	13.3	28.5	45.5	68.2	59.2	1.45	7.8	35.1	5.1
	$\epsilon \to 0$	3.3	14.1	35.6	13.0	28.1	44.8	66.5	60.5	1.42	7.6	35.4	4.7

TABLE 9
Influence of different inference formulations on MIT-States, UT-Zappos, and C-GQA. ✓ indicates that the model making inferences with these scores.

	Att/Obj	Comp	Complex	MIT-States				UT-Zappos				C-GQA			
				AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	AUC	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$
CW	✓			4.8	17.0	26.4	26.5	31.3	49.0	61.8	55.9	3.7	14.9	31.4	15.0
	✓	✓		5.3	18.1	31.3	25.9	36.9	52.6	67.7	70.2	4.6	17.3	34.0	17.0
	✓	✓	✓	8.0	23.6	35.2	30.1	41.4	55.3	68.0	72.3	5.4	18.9	35.4	18.6
OW	✓			2.4	10.4	32.5	13.0	27.6	42.4	67.0	55.2	0.70	4.5	31.1	3.1
	✓	✓		2.7	12.1	34.1	12.8	28.2	43.1	67.8	56.8	0.83	4.9	33.2	3.1
	✓	✓	✓	3.5	14.2	35.1	13.5	30.9	46.1	68.1	60.8	1.53	8.0	35.3	4.9

a positive incentive (37.1% vs 39.2% AUC for UT-Zappos). In summary, the enhancements from these two components in AGV are notably less pronounced when contrasted with the direct impacts of either complex space or OGA. Nevertheless, it is crucial to note that the feature decoupling facilitated by this foundational module is essential for training and inference in the complex space.

Impact of Affinity-based Pseudo Distribution. The APD module incorporates two primary types of pseudo-labels to construct distributions for unseen classes: 1) y_{k_j}, which represents the unseen composition most similar to the sample label y_j, and 2) compositions that share affinities with y_{k_j}. In this section, we evaluate the specific impacts of these two pseudo-labels and examine how the results vary when $\mathcal{L}_u$ is entirely removed. We manipulate these components by adjusting the values of ϵ and α. The results are presented in Tab. 8.

$\mathcal{L}_u$ establishes a pseudo-distribution for unseen classes through the generation of pseudo-labels, thereby mitigating bias between seen and unseen classes. When removing the $\mathcal{L}_u$ ($\alpha \to 0$), the results demonstrate an overall decrease in precision for unseen classes (*e.g.*, 27.5% vs 30.1% AUC on MIT-States). Compared to the results in Tab. 2 and Tab. 3, it decreases the overall performance across the seen and unseen classes, as evident in the decrease of AUC (7.4% vs 8.0%) and $\mathcal{A}^H$ (21.8% vs 23.6%). However, the enhancement of this loss function is not as significant in the OW-CZSL. We argue that this is due to the interference of the unseen class of infeasible compositions. Even though we can reduce these effects through hyper-parameter settings ϵ, they are still somewhat misleading to the model. This can also be reflected for the ϵ setting, where higher results are taken at $\epsilon \to 0$ in CW-CZSL. Whereas in OW-CZSL, higher results are taken from $\epsilon \to 1$. This suggests that the most similar unseen composition may not be feasible, which misrepresents the model alignment process.

Impact of Inference formulation. Our approach in Sec. 3.9 employs inference methods that incorporate multiple

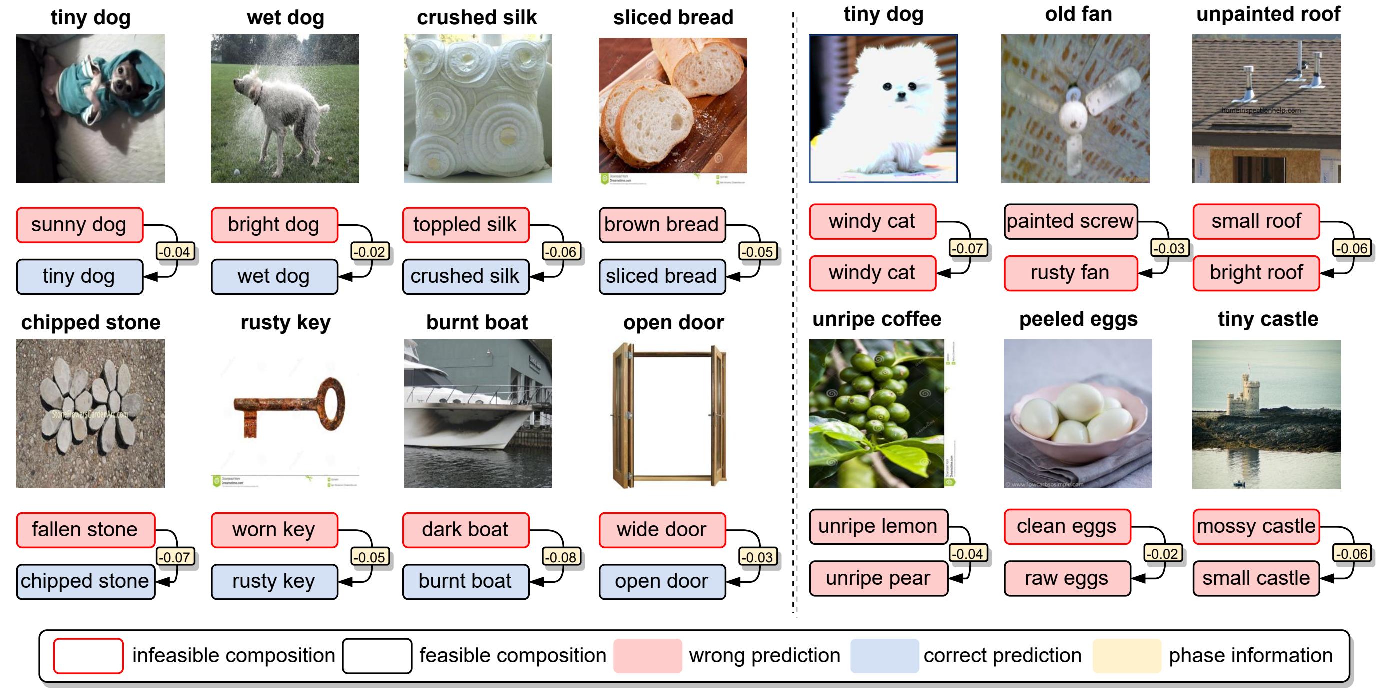


Fig. 4. Qualitative results of image retrieval on MIT-States for OW-CZSL. The prediction results for positive (left) and negative (right) cases are presented. The first row displays the results when phase information is not introduced, while the second row shows the results with phase information. Feasible compositions are marked with a black border, while infeasible ones are marked in red. The adjusted value of the phase information is indicated in the yellow box. The ground truth labels of the samples are provided above the respective images.

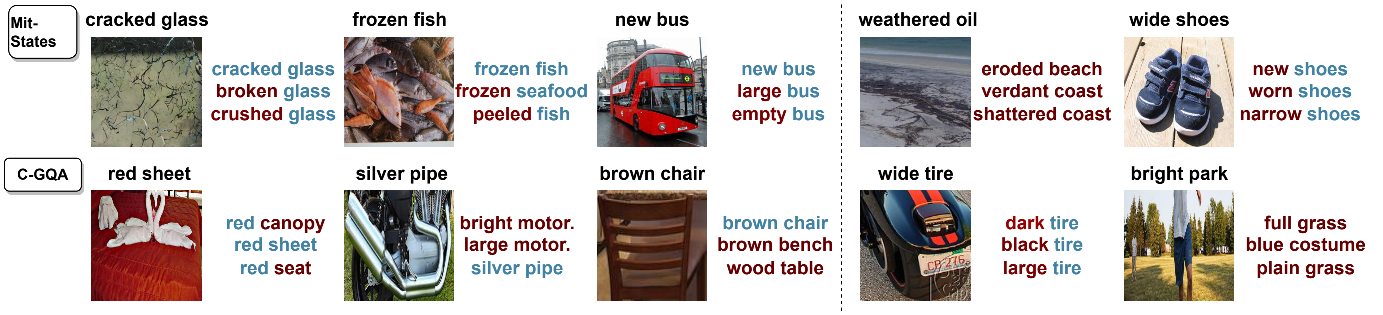


Fig. 5. Qualitative results of image retrieval on CW-CZSL, we show the results of top-3 predictions for images on MIT-States as well as on C-GQA, with correct ones marked in blue and incorrect ones in red. The ground truth labels of the samples are shown above the images.

scores. Here we validate the effectiveness of this approach. Specifically, we test the performance difference between the following similarity scores on the three benchmark datasets: (1) $\cos(\mathbf{z}^a, \mathbf{p}^a_j) \cdot \cos(\mathbf{z}^o, \mathbf{p}^o_j)$, denoted by *Att/Obj*, (2) $\cos(\mathbf{z}^r, \mathbf{p}^r_j)$, denoted by *Comp*, (3) $\rho_s(\mathbf{z}^c, \mathbf{p}^c_j) - \rho_p(\mathbf{z}^c, \mathbf{p}^c_j)$, denoted by *Complex*. The results are shown in Tab. 9.

In the context of *Att/Obj*, numerous prior studies have affirmed the positive impact of combining this component with *Comp* on enhancing CZSL task results [8], [11]. Departing from these approaches, we present complex-space inference outcomes in conjunction with this amalgamation, yielding the most significant enhancement among the three baselines (*e.g.*, achieving 3.5% AUC for MIT-States in OW-CZSL). Both *Att/Obj* and *Comp* represent inferences in real space, while *Complex* denotes the outcome of an inference in complex space. This suggests that inference in complex space are potential for synergistic combination with results in real space to augment information. Additionally, the results obtained using only *Att/Obj* also outperform certain recent methods, such as OADis [8], we attribute this mainly to the introduction of the AGV and OGA modules.

4.6 Cross-Dataset Results

In accordance with [9], we conduct an experiment involving cross-data analysis. In the aforementioned experiments (Sec. 4.4 and Sec. 4.5), our training and test data are sourced consistently, characterized by the same style and a similar range of categories. However, achieving such consistency in a real-world scenario is challenging, and it is desirable for the model to possess the ability to categorize even when confronted with entirely inconsistent input samples post-training. Specifically, we evaluate the model's performance in scenarios where the distributions of training and test data do not align. To implement this, we utilize two datasets, MIT-States [51] and C-GQA [5], and subject the model to

TABLE 10
Cross-dataset results on MIT-States and C-GQA in OW-CZSL. We evaluate the seen accuracy ($\mathcal{A}^S$), unseen accuracy ($\mathcal{A}^U$), and harmonic mean ($\mathcal{A}^H$) for comparison purposes.

Training	MIT-States			C-GQA		
Test	C-GQA			MIT-States		
	$\mathcal{A}^S$	$\mathcal{A}^U$	$\mathcal{A}^H$	$\mathcal{A}^S$	$\mathcal{A}^U$	$\mathcal{A}^H$
SymNet [6]	6.5	0.93	0.83	0.44	0.21	0.10
CGE [5]	6.3	1.1	1.0	0.38	0.21	0.13
CompCos [10]	6.3	2.5	1.5	0.59	0.49	0.17
Co-CGE$^{\text{CW}}$ [9]	6.2	1.1	1.0	0.91	0.33	0.15
Co-CGE [9]	5.5	3.2	1.6	0.80	0.31	0.19
IMAX*	6.4	3.5	1.8	0.95	0.44	0.23

testing on a different dataset after it has been fully trained on either one of them.

Since it is ineffective to partition the validation set in this setting, we directly use the weights of the trained model in Sec. 4.4 for testing here, the results are shown in Tab. 10. For a fair comparison with these methods, IMAX uses R-18 as the backbone, the same as the compared methods. The results are shown for the OW-CZSL. As a result of the change in the data distribution, we can observe a substantial dip in results for all methods. For example, the $\mathcal{A}^H$ for CGE drops from 29.6% to 6.3%. This is since there are also a large number of compositions in the seen compositions that have not been seen by the model in this scenario.

In our experimental analysis, when training our approach on the MIT-States dataset and evaluating it on C-GQA, we observe that our method demonstrates robust generalization to unseen classes, surpassing Co-CGE by $+0.3\%$. However, our model's discriminative capability for seen classes falls slightly short by -0.1% compared to SymNet. Nonetheless, when considering the harmonic accuracy, which provides a comprehensive measure of overall performance, our approach took a complete lead, which outperforms Co-CGE by 0.2%. When training on the C-GQA and evaluating IMAX on MIT-States, we can observe similar trends. IMAX outperforms all the methods in $\mathcal{A}^S$ ($+0.04\%$ compared to Co-CGE$^{\text{CW}}$) but is slightly lower than CompCos in $\mathcal{A}^U$ (-0.05%), while it continues to lead in the most important metric $\mathcal{A}^H$ ($+0.04\%$). The aforementioned results demonstrate that our method exhibits versatility across multiple datasets, rather than being limited to the OW-CZSL task with a single data distribution, which is important for applications in real-world scenarios.

4.7 Qualitative Results

How complex phase information is involved in classification. Empirical evidence presented in Sec. 4.5 demonstrates the efficacy of phase information in CW-CZSL and OW-CZSL by facilitating an improved understanding of intra-constituent contextual relationships through dependency construction. In this section, we present the composition of predicted scores generated by IMAX for different samples. Fig. 4 illustrates both the predictions without phase information and those with phase information when using ViT-B as the backbone. Furthermore, we report the values of phase information, thus illustrating how the model leverages complex phase information to discern infeasible compositions and comprehend the relationship between attributes and objects.

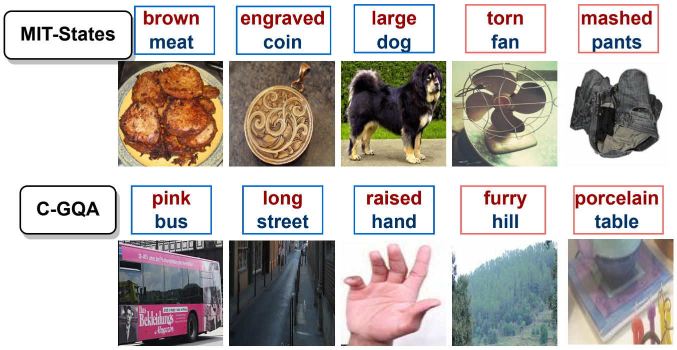


Fig. 6. Qualitative results of text retrieval. We use feasible and infeasible composition semantics for image retrieval, with feasible ones labeled with a blue border and infeasible ones in red. Images are top-1 retrieval results for each text.

Fig. 4 summarizes several significant trends. First, direct prediction of compositions in real space is prone to infeasibility. For instance, consider the first sample in the top row where the `dog` is exposed to light sources. In this case, the real-space approach directly predicts `sunny`, disregarding its association with the `dog`. In contrast, incorporating complex phase information reduces the likelihood of predicting such compositions, leading to accurate inferences. Numerous instances of the same nature can be observed, particularly concerning attributes, whereby real-space predictions yield outcomes such as `bright`, `worn`, `dark`, `wide`, `brown`, and so on. While there may be some plausibility regarding attributes, real-space predictions disregard the contextual relationship they share with associated objects (`dog`, `key`, `boat`, `door`, `bread`).

On the right side of Fig. 4, the samples are displayed where errors persist despite the incorporation of phase information. These errors primarily arise in the following situations: 1) Inability to accurately classify objects. For instance, when presented with a sample of `unripe coffee`, our model struggles to properly categorize it due to the significant bias introduced by the concept of being unripe. This limitation arises from data constraints, impeding our model's ability to generalize effectively in this context. 2) Incorrect attributes and object localization. Take the sample of `old fan`, for instance. The inclusion of phase information may lead to an even less infeasible outcome. Despite the presence of an approximate `rusty` component in the figure, it does not manifest in the `fan` itself, and as a result, our model fails to establish the connection between this component and the object accurately.

Image and text retrieval. CZSL has a prospective multi-label property. Each object may exist in multiple attributes; for instance, a *tree* may be both *tall* and *green*. Therefore, we perform an image-to-text retrieval in CW-CZSL when using ViT-B as the backbone, and present the top three predictions for various samples from various datasets in Fig. 5.

Fig. 5 demonstrates that the accurate samples on the left side indicate that our model possesses the capability to go beyond matching correct labels and instead generalize the visual-semantic relationship. For instance, in the case of the `cracked glass` sample, our model provides three predic-

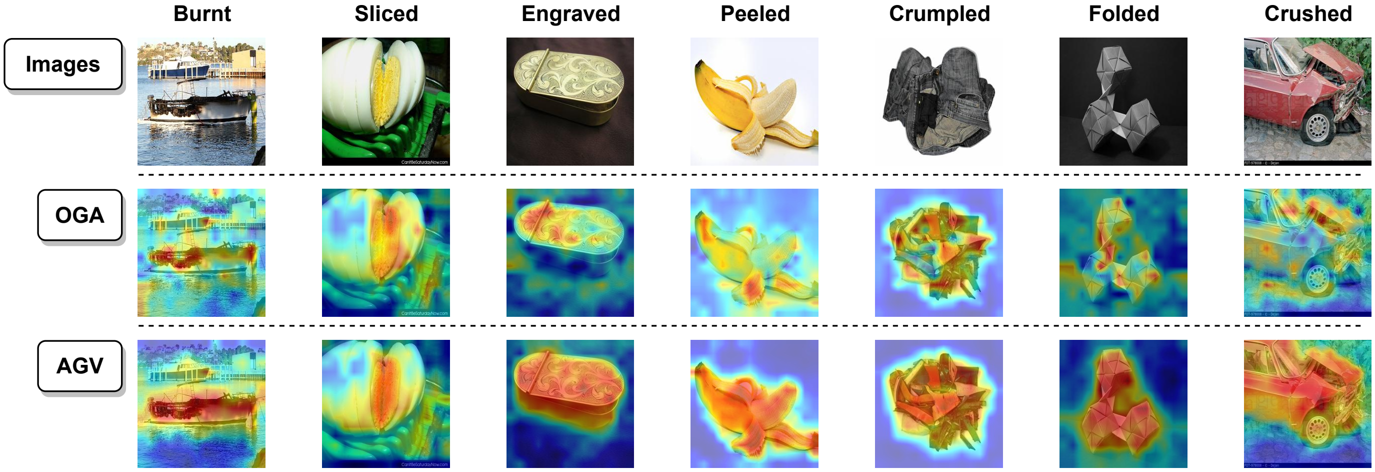


Fig. 7. Visualization of AGV and OGA on MIT-States about attribute extraction. The top row displays the original image, with its associated attributes labeled above. The second row depicts the region of interest for the attributes within the OGA module, whereas the third row showcases the region of interest for the attributes within the AGV module.

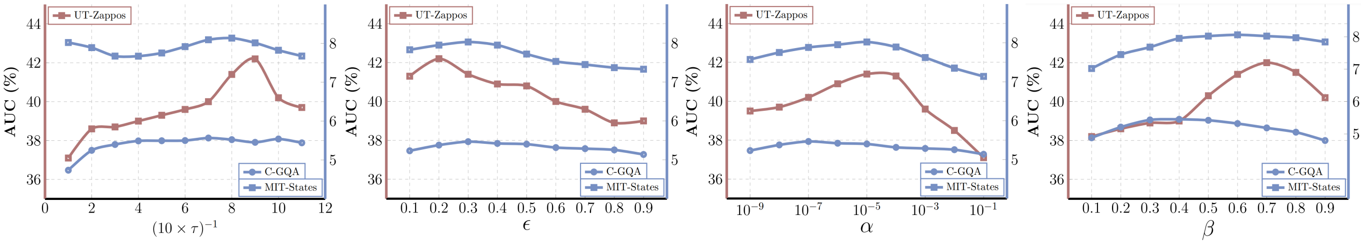


Fig. 8. Hyper-parameter analysis in CW-CZSL. We report the AUC under different hyper-parameter settings for UT-Zappos (left Y-axis, red) and for both C-GQA and MIT-States (right Y-axis, blue).

tions: `cracked glass`, `broken glass`, and `crushed glass`. Considering the semantic similarity among these attributes, we deem all these outcomes as plausible. Considering the subjective nature of dataset labeling, we find it justifiable to provide prediction results for either `cracked` or `broken`, as this sample essentially exhibits a multi-labeled nature. Furthermore, we eagerly anticipate future proposals for a multi-labeled CZSL dataset.

The inaccurate prediction results on the right-hand side reveal certain limitations of our model. For instance, in the scenario involving `weathered oil`, our model erroneously emphasizes the ocean-related features like `coast` and `beach`, disregarding the primary object, `oil`, thereby leading to a complete misjudgment of the attribute. This indicates that object localization in AGV remains inadequate, particularly in handling objects like `oil` that lack a clearly defined external structure.

Moreover, we conduct experiments on text-to-image retrieval. We employ IMAX to compute the images most similar to those provided with a semantic description. We present the retrieval results for ten different texts in MIT-States and C-GQA, including the results for both feasible (blue box) and infeasible (red box) compositions. Fig. 6 displays the results, with our specific focus being on the observation of retrieving infeasible compositions, such as `torn fan` and `mashed pants`, among others. The model is capable of retrieving sensible images, which demonstrates its ability to attain a comprehensive comprehension of comparable semantics. In cases of ambiguous semantics, the model considers the object as the primary entity and seeks a comparable attribute as a replacement.

Visual analysis for attribute extraction. Our attribute extraction primarily involves fusing the outputs of the AGV and OGA modules. The AGV module is employed to localize items to be classified. Meanwhile, the OGA module models the visual bias imposed by the attributes by capturing variations from the object prototype. As a comparison, we show the CAM [65] heat map for the AGV module as well as the OGA module when extracting attributes, respectively, shown in Fig. 7.

In the figure, we can observe two clear trends: 1) OGA tends to acquire the localized region with the largest visual bias, whereas AGV contains object regions that are independent of the attribute, due to the depth entanglement within the composition. 2) OGA modules are prone to cover their background noise due to the backbone is not fine-tuned, whereas AGV modules usually only cover their target body. These two phenomena can explain our fusion of these two modules as attribute extraction, *i.e.*, OGA provides the most significant attribute regions, while AGV is responsible for the approximate location of the attributes.

4.8 Hyper-Parameter Analysis

In this section, our focus lies in evaluating the impact of four hyper-parameters on CW-CZSL, namely: 1) the temperature coefficient τ, 2) the weighting adjustment factor ϵ, 3) the loss weighting coefficient α, and 4) the scores fusion weights. The results, presented in Fig. 8, reveal that on MIT-States, UT-Zappos, and C-GQA, the reciprocal of $\frac{1}{\tau}$ attains its peak value at 90, 80 and 70, respectively. Conversely, ϵ reaches its maximum value of 0.2, 0.3 and 0.3, surpassing the result we reported in Tab. 2. These differences primarily arise from the disparity between the validation set and the test set. Concerning α, it reaches its peak at 10^{-5}, 10^{-5} and 10^{-7}. Due to the absence of data for unseen classes, excessively large values of α may introduce bias in the predictions for seen classes. The peak value of β occurs at 0.7, 0.6, and 0.3, indicating that a moderate fusion of prediction results from multiple modules yields benefits for the overall inference. On the whole, the fluctuations resulting from these hyperparameters fall within an acceptable range, and even if the optimal values are not chosen, our method still outperforms the results presented in Tab. 2.

5 Conclusion

In this paper, we address the CZSL problem, aiming to recognize unseen compositions from seen attribute-object pairs. Our approach tackles both CW-CZSL and OW-CZSL by introducing an imaginary-connected embedding structure that integrates semantic dependencies and primitives. Specifically, we decompose visual features into the complex space using AGV and OGA, with attributes as an additional imaginary unit. We also introduce a pseudo-distribution of unseen classes to enhance the model's generalization to unseen classes. These improvements enable our model to leverage phase information for higher-dimensional similarity calculations, incorporating intra-compositional dependencies in CW-CZSL and modeling compositional feasibility in OW-CZSL. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches, though it also reveals limitations, particularly in misidentifying attributes and objects in samples with multiple entities. This suggests that IMAX requires a deeper semantic understanding of the samples.

References

[1] B. M. Lake, "Towards more human-like concept learning in machines: Compositionality, causality, and learning-to-learn," Ph.D. dissertation, Massachusetts Institute of Technology, 2014.

[2] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, p. e253, 2017.

[3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," in *NAACL*, 2016, pp. 1545–1554.

[4] I. Misra, A. Gupta, and M. Hebert, "From red wine to red tomato: Composition with context," in *CVPR*, 2017, pp. 1792–1801.

[5] M. F. Naeem, Y. Xian, F. Tombari, and Z. Akata, "Learning graph embeddings for compositional zero-shot learning," in *CVPR*, 2021, pp. 953–962.

[6] Y.-L. Li, Y. Xu, X. Mao, and C. Lu, "Symmetry and group in attribute-object compositions," in *CVPR*, 2020, pp. 11 316–11 325.

[7] X. Li, X. Yang, K. Wei, C. Deng, and M. Yang, "Siamese contrastive embedding network for compositional zero-shot learning," in *CVPR*, 2022, pp. 9326–9335.

[8] N. Saini, K. Pham, and A. Shrivastava, "Disentangling visual embeddings for attributes and objects," in *CVPR*, 2022, pp. 13 658–13 667.

[9] M. Mancini, M. F. Naeem, Y. Xian, and Z. Akata, "Learning graph embeddings for open world compositional zero-shot learning," *IEEE TPAMI*, 2022.

[10] ——, "Open world compositional zero-shot learning," in *CVPR*, 2021, pp. 5222–5230.

[11] S. Hao, K. Han, and K.-Y. K. Wong, "Learning attention as disentangler for compositional zero-shot learning," in *CVPR*, 2023, pp. 15 315–15 324.

[12] Q. Wang, L. Liu, C. Jing, H. Chen, G. Liang, P. Wang, and C. Shen, "Learning conditional attributes for compositional zero-shot learning," in *CVPR*, 2023, pp. 11 197–11 206.

[13] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.

[14] J. Von Neumann, *Mathematical foundations of quantum mechanics: New edition*. Princeton university press, 2018, vol. 53.

[15] H. Kim, J. Lee, S. Park, and K. Sohn, "Hierarchical visual primitive experts for compositional zero-shot learning," in *ICCV*, 2023, pp. 5675–5685.

[16] D. D. Hoffman and W. A. Richards, "Parts of recognition," *Cognition*, vol. 18, no. 1-3, pp. 65–96, 1984.

[17] I. Biederman, "Recognition-by-components: a theory of human image understanding." *Psychological review*, vol. 94, no. 2, p. 115, 1987.

[18] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *ICCV*, 2017, pp. 3018–3027.

[19] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, vol. 30, 2017.

[20] J. He, A. Kortylewski, and A. Yuille, "Corl: Compositional representation learning for few-shot classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3890–3899.

[21] S. Pratt, I. Covert, R. Liu, and A. Farhadi, "What does a platypus look like? generating customized prompts for zero-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 691–15 701.

[22] S. Menon and C. Vondrick, "Visual classification via description from large language models," *arXiv preprint arXiv:2210.07183*, 2022.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[24] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE TPAMI*, 2022.

[25] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *CVPR*, 2013, pp. 819–826.

[26] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE TPAMI*, vol. 36, no. 3, pp. 453–465, 2013.

[27] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *NeurIPS*, 2013, p. 2121–2129.

[28] T. Paz-Argaman, Y. Atzmon, G. Chechik, and R. Tsarfaty, "Zest: Zero-shot learning from text descriptions using textual similarity and visual summarization," *arXiv preprint arXiv:2010.03276*, 2020.

[29] D. Chen, Y. Shen, H. Zhang, and P. H. Torr, "Zero-shot logit adjustment," *arXiv preprint arXiv:2204.11822*, 2022.

[30] Z. Xu, G. Wang, Y. Wong, and M. S. Kankanhalli, "Relation-aware compositional zero-shot learning for attribute-object pair recognition," *IEEE TMM*, vol. 24, pp. 3652–3664, 2021.

[31] W. Xu, Y. Xian, J. Wang, B. Schiele, and Z. Akata, "Attribute prototype network for zero-shot learning," in *NeurIPS*, 2020, pp. 21 969–21 980.

[32] M. Elhoseiny, Y. Zhu, H. Zhang, and A. Elgammal, "Link the head to the" beak": Zero shot learning from noisy text description at part precision," in *CVPR*, 2017, pp. 5640–5649.

[33] R. Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[34] C.-Y. Chen and K. Grauman, "Inferring analogous attributes," in *CVPR*, 2014, pp. 200–207.

[35] M. Yang, C. Deng, J. Yan, X. Liu, and D. Tao, "Learning unseen concepts via hierarchical decomposition and composition," in *CVPR*, 2020, pp. 10 248–10 256.

[36] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *ECCV*. Springer, 2016, pp. 852–869.

[37] Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik, "A causal view of compositional zero-shot recognition," in *NeurIPS*, 2020, pp. 1462–1473.

[38] M. Yang, C. Xu, A. Wu, and C. Deng, "A decomposable causal view of compositional zero-shot learning," *IEEE TMM*, 2022.

[39] S. Karthik, M. Mancini, and Z. Akata, "Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning," in *CVPR*, 2022, pp. 9336–9345.

[40] N. V. Nayak, P. Yu, and S. H. Bach, "Learning to compose soft prompts for compositional zero-shot learning," *arXiv preprint arXiv:2204.03574*, 2022.

[41] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[42] X. Lu, S. Guo, Z. Liu, and J. Guo, "Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning," in *CVPR*, 2023, pp. 23 560–23 569.

[43] W. Bao, L. Chen, H. Huang, and Y. Kong, "Prompting language-informed distribution for compositional zero-shot learning," *arXiv preprint arXiv:2305.14428*, 2023.

[44] T. Nagarajan and K. Grauman, "Attributes as operators: factorizing unseen attribute-object compositions," in *ECCV*, 2018, pp. 169–185.

[45] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," *arXiv preprint arXiv:1902.10197*, 2019.

[46] Y. Fu, Y. Xie, Y. Fu, and Y.-G. Jiang, "Styleadv: Meta style adversarial training for cross-domain few-shot learning," in *CVPR*, 2023, pp. 24 575–24 584.

[47] S. Lin, Z. Zhang, Z. Huang, Y. Lu, C. Lan, P. Chu, Q. You, J. Wang, Z. Liu, A. Parulkar *et al.*, "Deep frequency filtering for domain generalization," in *CVPR*, 2023, pp. 11 797–11 807.

[48] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," in *NeurIPS*, 2018.

[49] S. Lee, J. Bae, and H. Y. Kim, "Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization," in *CVPR*, 2023, pp. 11 776–11 785.

[50] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE TPAMI*, vol. 41, no. 9, pp. 2251–2265, 2018.

[51] P. Isola, J. J. Lim, and E. H. Adelson, "Discovering states and transformations in image collections," in *CVPR*, 2015, pp. 1383–1391.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[54] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[55] J. Xie, J. Xiang, J. Chen, X. Hou, X. Zhao, and L. Shen, "C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation," in *CVPR*, 2022, pp. 989–998.

[56] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[57] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017, pp. 1501–1510.

[58] S. Liu, M. Long, J. Wang, and M. I. Jordan, "Generalized zero-shot learning with deep calibration network," in *NeurIPS*, 2018.

[59] H. Zhang, J. Liu, Y. Yao, and Y. Long, "Pseudo distribution on unseen classes for generalized zero shot learning," *Pattern Recognition Letters*, vol. 135, pp. 451–458, 2020.

[60] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *CVPR*, 2014, pp. 192–199.

[61] S. Purushwalkam, M. Nickel, A. Gupta, and M. Ranzato, "Task-driven modular networks for zero-shot compositional learning," in *ICCV*, 2019, pp. 3593–3602.

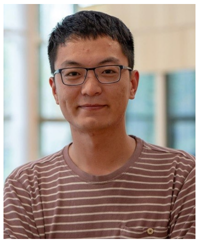
[62] X. Wang, F. Yu, T. Darrell, and J. E. Gonzalez, "Task-aware feature generation for zero-shot compositional learning," *arXiv preprint arXiv:1906.04854*, 2019.

[63] Y. Li, Z. Liu, S. Jha, and L. Yao, "Distilled reverse attention network for open-world compositional zero-shot learning," in *ICCV*, 2023, pp. 1782–1791.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[65] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.

Chenyi Jiang received the B.S. degree in Mathematics from Fuzhou University, Fuzhou, China, in 2021, and is currently working toward the PhD degree in the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His current research interests include Zero-shot Learning and Few-shot Learning.



Shidong Wang is a Lecturer in Data Engineering & AI, School of Computing, Newcastle University, UK. He received his PhD degree from the School of Computing Sciences, University of East Anglia (UEA), UK, in 2021. His research spans a breadth of domains including computer vision, deep learning, remote sensing, and environmental science, and he publishes in top-tier journals and conferences such as IEEE TPAMI, TIP, IJCV, ISPRS, AAAI and ACM MM.

Yang Long is an Assistant Professor in the Department of Computer Science, Durham University. He is also an MRC Innovation Fellow aiming to design scalable AI solutions for large-scale healthcare applications. His research background is in the highly interdisciplinary field of Computer Vision and Machine Learning. While he is passionate about unveiling the black-box of AI brain and transferring the knowledge to seek Scalable, Interactable, Interpretable, and sustainable solutions for other disciplinary researches. He has authored/co-authored 20+ top-tier papers in refereed journals/conferences such as IEEE TPAMI, TIP, CVPR, AAAI, and ACM MM.

Zechao Li Zechao Li is currently a Professor at the Nanjing University of Science and Technology. He received the Bachelor's degree from University of Science and Technology of China (USTC) in 2008 and the Doctor degree from the Institute of Automation, Chinese Academy of Sciences in 2013. He has authored over 90 papers in top-tier journals and conferences. His research interests include multimedia analysis, object detection, semantic segmentation, few-shot learning, etc. He has served as an Associate Editor for IEEE TPAMI, TNNLS, JCST, etc.

Haofeng Zhang is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. He received the B.Eng. degree and the Ph.D. degree in 2003 and 2007 respectively from School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, China. From Dec. 2016 to Dec. 2017, He was an academic visitor at University of East Anglia, Norwich, UK. His research interests include computer vision and mobile robot.

Ling Shao (Fellow, IEEE) is a distinguished professor with the UCAS-Terminus AI Lab, University of Chinese Academy of Sciences, Beijing, China. He was the founding CEO and chief scientist of the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. He was also the initiator, founding provost and EVP of MBZUAI, UAE. His research interests include generative AI, vision and language, and AI for healthcare. He is a fellow of the IEEE, the IAPR, the BCS and the IET.