Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Expanding the Category of Classifiers with LLM Supervision

Derui Lyu, **Xiangyu Wang***, **Taiyu Ban**, **Lyuzhou Chen**, **Xiren Zhou**, **Huanhuan Chen***
University of Science and Technology of China
{drlv, banty, clz31415}@mail.ustc.edu.cn, {sa312, zhou0612, hchen}@ustc.edu.cn

Abstract

Zero-shot learning has shown significant potential for creating cost-effective and flexible systems to expand classifiers to new categories. However, existing methods still rely on manually created attributes designed by domain experts. Motivated by the widespread success of large language models (LLMs), we introduce an LLM-driven framework for class-incremental learning that removes the need for human intervention, termed Classifier Expansion with Multi-vIew LLM knowledge (CEMIL). In CEMIL, an LLM agent autonomously generates detailed textual multi-view descriptions for unseen classes, offering richer and more flexible class representations than traditional expert-constructed vectorized attributes. These LLM-derived textual descriptions are integrated through a contextual filtering attention mechanism to produce discriminative class embeddings. Subsequently, a weight injection module maps the class embeddings to classifier weights, enabling seamless expansion to new classes. Experimental results show that CEMIL outperforms existing methods using expert-constructed attributes, demonstrating its effectiveness for fully automated classifier expansion without human participation.

1 Introduction

As the field evolves and data grows, new categories are often discovered or redefined, creating shifting demands for existing classifiers. Consequently, classifiers need to expand to accommodate emerging unseen categories, termed the *classifier expansion* task. Re-training with an expanded class set serves as a traditional solution, but requires significant image data collection and repetitive training, costly in many contexts.

The advent of zero-shot learning (ZSL) has inspired a zero-shot classifier expansion paradigm, which relies solely on images of seen classes [Xian *et al.*, 2017; Wei *et al.*, 2021; Xu *et al.*, 2020]. Recent studies have further explored classifier expansion in completely image-free settings [Christensen *et al.*, 2023; Yun *et al.*, 2023]. These methods expand existing classifiers to recognize unseen classes by aligning expert-constructed attribute features with the classifier, thus integrating new classes into the visual embedding space without the need for any images. However, while these approaches reduce the image data requirements compared to re-training methods, they still depend on expert input, as attributes must be carefully designed and annotated, which keeps human costs involved [Xian *et al.*, 2017]. We summarize the category expansion approaches, as illustrated in Figure 1.

*Corresponding authors.

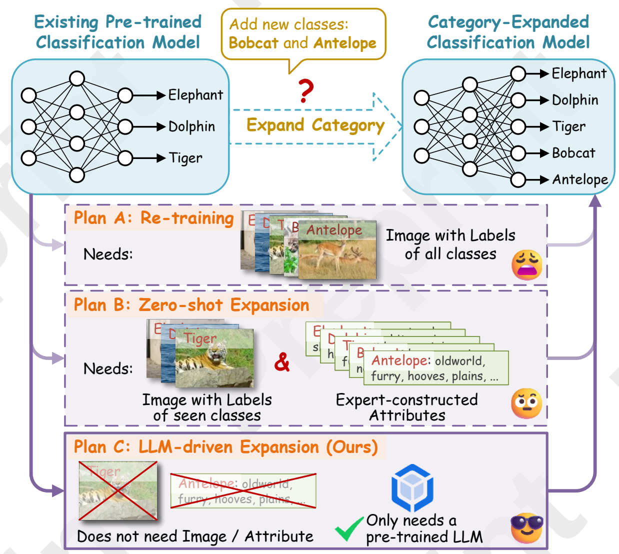


Figure 1: Overview of category expansion approaches. Expanding a classifier to new categories can be done via: *A. Retraining*: Requires labeled images for all classes. *B. Zero-shot Expansion*: Needs seen-class images and expert-constructed attributes. *C. CEMIL*: Uses a pre-trained LLM, removing the need for images or attributes.

Inspired by the success of LLMs in reducing manual effort across various domains, we explore their potential to eliminate human dependency in classifier expansion. For this task, LLMs excel at generating detailed textual descriptions for a given class, offering valuable insights for "teaching" classifiers to recognize new labels. However, three key challenges remain. First, the descriptions generated by LLMs may not consistently maintain high quality for different classes, and often fail to offer the depth of information required to fully

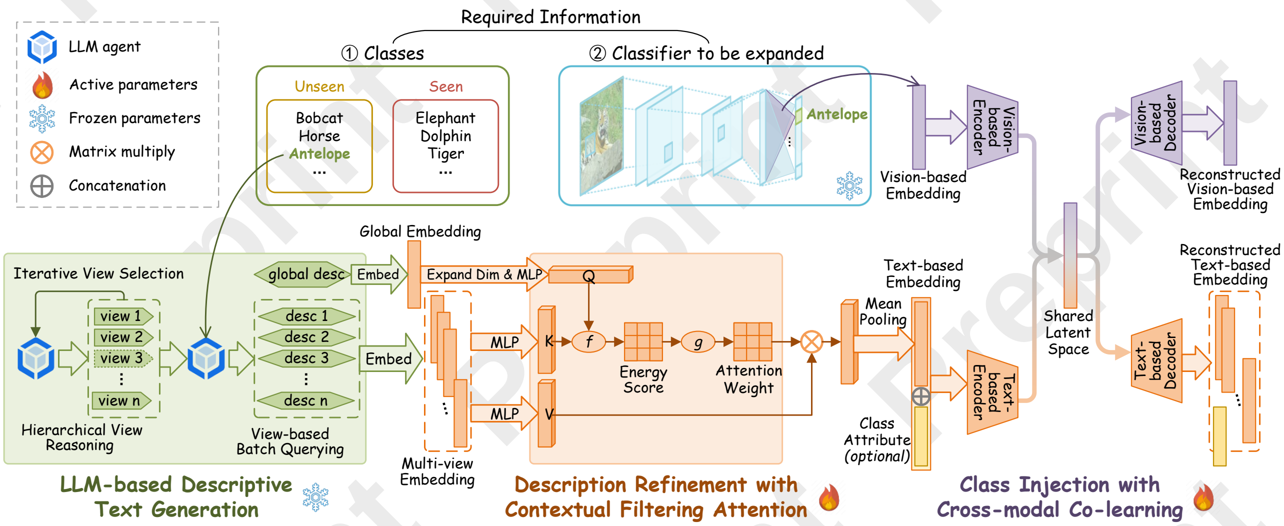


Figure 2: Framework of the proposed CEMIL approach. CEMIL begins with a set of class names and generates robust multi-view features through the LLM (DTG module). These features are then refined using a contextual filtering attention mechanism (CFA module), and used to supervise the classifier's weights via a cross-modal co-learning process, facilitated by a dual-autoencoder (CICC module). The predicted weights for unseen classes can be directly injected into the output layer of the classifier, enabling its seamless expansion. Since the embedding model can be implemented with an open-source LLM, the entire workflow is fully automated, requiring no human involvement.

capture the characteristics of new classes. Second, the natural language output from LLMs frequently includes redundant or irrelevant details, requiring a refinement process to distill only the most essential features. Finally, the class features embedded in LLM-generated textual descriptions and those in the classifier are represented in different modalities, with significant differences in how these class representations are captured, making it complex to align them effectively.

To address these issues, this paper presents an LLM-driven classifier expansion framework that removes the need for manually designed attributes, termed **C**lassifier **E**xpansion with **M**ulti-v**I**ew **L**LM Knowledge (CEMIL). CEMIL consists of three sequential modules: LLM-based Descriptive Text Generation (DTG), Description Refinement with Contextual Filtering Attention (CFA), and Class Injection with Cross-modal Co-learning (CICC). The DTG module introduces an LLM reasoning flow to automatically derive comprehensive, hierarchical descriptions for classes. These multi-level descriptions are then embedded into refined representations by the CFA module, which employs high-level descriptions as contextual attention to filter out less relevant information. Finally, the CICC module aligns the refined embeddings with the classifier weights, enabling seamless cross-modal expansion of the classifier to new classes. Extensive experiments demonstrate that CEMIL, which operates solely with an LLM, consistently outperforms state-of-the-art ZSL methods that rely on expert-constructed attributes, across a variety of ZSL datasets. Figure 2 illustrates the framework of the CEMIL method. We summarize our contributions as follows:

- We propose a novel LLM-driven paradigm that enables the category expansion of existing classifiers without requiring any images or additional data. To the best of our knowledge, we are the first to use LLMs as a source of supervision for human-free classifier expansion.
- We introduce CEMIL, a novel approach that: i) directs LLMs to deliver comprehensive and robust multi-view information, ii) extracts discriminative features from weak and noisy texts, and iii) effectively guides classifier expansion. This design stimulates and fully exploits the power of LLMs for cross-modal supervision.
- CEMIL achieves state-of-the-art performance on multiple image-free ZSL benchmarks, consistently demonstrating superior stability across various configurations. It holds the potential to reshape traditional ZSL tasks by introducing a new paradigm of LLM-based supervision.

2 Related Work

2.1 Category Expansion for Existing Classifiers

The task of classifier expansion involves extending an existing classifier to recognize new, unseen classes that were not defined during its training. As the application scenarios for classifiers evolve, this task has become increasingly critical and commonplace. However, the challenge lies in achieving effective expansion with minimal or cost-effective supervision, a key area of ongoing research primarily explored within the field of class-incremental learning (CIL) [Mittal *et al.*, 2021; Wang *et al.*, 2024]. Previous studies have extensively explored CIL in few-shot settings [Hersche *et al.*, 2022; Zhou *et al.*, 2024]. Unlike traditional re-training methods, few-shot CIL aims to expand classifiers using a smaller number of images from new classes, achieving notable success. As the available data continues to diminish in zero-shot settings, no image data from the new classes is available, and the ZSL models rely on additional side information to distinguish between classes. This information typically comes from expert-constructed class attributes [Li *et al.*, 2023] or external textual sources, such as Wikipedia [Zhu *et al.*, 2018].

The methods mentioned above typically rely on image data from seen classes, which is often unavailable in real-world scenarios. Consequently, some approaches explore classifier expansion in image-free settings by predicting and directly injecting the output layer parameters of new classes into the existing classifier, using expert-defined attributes to bridge the gap between seen and unseen classes. A straightforward approach involves using a multi-layer perceptron (MLP) to map visual classifier weights to attributes, which performs well in ZSL but tends to degrade in generalized settings. To address this limitation, Christensen et al. [Christensen *et al.*, 2023] propose specific autoencoders for both attribute and weight spaces, regularizing the attribute-to-weight mapping and achieving strong performance on the image-free generalized ZSL task. Additionally, building on this image-free idea, other methods [Norouzi *et al.*, 2014; Mensink *et al.*, 2014; Akyürek *et al.*, 2022; Xu *et al.*, 2022] can also be adapted to expand classifiers in the image-free zero-shot setting.

2.2 LLM-driven Cross-modal Supervision

Zero-shot learning tasks require the use of auxiliary information to predict unseen categories. Most traditional approaches rely on expert-crafted attributes, such as manually designed numerical features [Lampert *et al.*, 2013; Xian *et al.*, 2017]. While these features are accurate and discriminative, they require manually defining attribute names and expert annotations for attribute values. This process is resource-intensive, particularly in fine-grained datasets [Wah *et al.*, 2011]. Some methods use text descriptions from external sources like Wikipedia [Qiao *et al.*, 2016; Zhu *et al.*, 2018] or summary documents [Naeem *et al.*, 2024] to reduce expert dependence, but this supervision is still manually constructed and limited in flexibility and information due to its disconnection from the ZSL task.

Recently, the rise of LLMs has shown great potential for enhancing few-shot or zero-shot learning tasks [Guo *et al.*, 2023; Ban *et al.*, 2025]. As a flexible knowledge source with advantages in both the quality and quantity of supervision, previous works have explored using LLMs for cost-effective knowledge expansion [Li *et al.*, 2024; Wu *et al.*, 2024]. An early work is I2MVFormer [Naeem *et al.*, 2023], which uses multi-view prompting to encode LLM-supervised semantic embeddings for zero-shot image classification, while Adapt-CLIPZS [Saha *et al.*, 2024] extends this to pre-trained vision-language models (VLMs), achieving improved results. However, these methods still rely on image samples for training and fine-tuning. As an improvement, Liu et al. [Liu *et al.*, 2024] explore training models solely on textual data by developing a cross-modal classifier with LLMs and mapping it to the visual modality, achieving zero-shot multi-label recognition with pre-trained VLMs. Despite these advances, these methods still heavily depend on pre-trained large-scale image encoders, such as CLIP, making them less suitable for scenarios involving existing image classifiers [Zhao *et al.*, 2024].

In contrast, this paper uses only the cheapest LLM-based semantic knowledge to extend existing classifiers, without any image data or VLMs trained on large amounts of images. This approach allows for classifier expansion with minimal annotation costs, data dependency, and domain bias.

3 Methodology

The entire CEMIL process consists of the following three stages: **1) Text Generation**: Generate descriptive texts using a pre-trained LLM as the agent (DTG module). **2) Description Refinement**: Distill and integrate global and multi-view local descriptions using contextual filtering attention (CFA module). **3) Class Injection**: Learn and inject classifier weights into the existing model through a cross-modal co-learning framework (CICC module). In this section, we first define the task and then introduce each stage in detail.

3.1 Task Formulation

We address the challenge of expanding a pre-trained classifier for unseen classes, without any image or attribute. Given a pre-trained classifier Φ: $\mathcal{X} \rightarrow C_s$, where $\mathcal{X}$ represents the image space and C_s denotes the set of seen classes, the objective of the *Classifier Expansion* is to expand the classifier's capability to classify a new class set C and get an expanded classifier $\hat{\Phi}$: $\mathcal{X} \rightarrow C$. In the standard ZSL setting, the target set C is entirely disjoint from the seen classes, *i.e.* $C \cap C_s = \emptyset$. In the generalized ZSL setting, the target set C includes both the seen and unseen classes, *i.e.* $C = C_s \cup C_u$. Previous ZSL works typically use an expert-based class attribute matrix $A \in \mathbb{R}^{m \times d_a}$, while this paper addresses the task using *only* an LLM $\mathcal{M}$, and treat A as *optional*.

Without loss of generality, the classifier Φ can be divided into a feature extractor ω and a classification layer ψ, *i.e.*, $\Phi = \psi(\omega(\cdot))$. For a neural network classifier, the classification layer ψ is parameterized by a matrix, *i.e.*, $\psi \in \mathbb{R}^{d_v \times |C_s|}$, where d_v is the embedding dimension of the classifier and the output dimension of ω. Since the task operates under the image-free setting, the goal is indeed to obtain an expanded classification layer parameter $\hat{\psi} \in \mathbb{R}^{d_v \times m}$.

3.2 LLM-based Descriptive Text Generation

We design a workflow for automatically generating descriptive texts from scratch, using an LLM as the agent. Without prior knowledge of the classifier, we first reason a comprehensive and effective set of views. Based on this, we employ a view-based batch querying strategy to generate descriptive and comparable multi-view texts for each class.

Structured Hierarchical View Reasoning
View generation can be complex for LLMs, as class identification often requires fine-grained views, which can lead to omissions or overlap, especially when the view number is large. Considering the limited capacity of LLMs, we aim to decompose the view generation task into smaller subtasks, enabling the LLM to tackle them in a systematic manner. Specifically, we guide the LLM to first generate coarse-grained perspectives (e.g., physical traits, behavioral traits) and then refine them into fine-grained details (e.g., fur color, feeding habits). This hierarchical divide-and-conquer approach reduces the LLM's capacity demand in a single pass.

To further enhance the quality, we prompt the LLM to act as a domain expert, leveraging its specialized knowledge, and provide structured examples to harness its in-context learning ability for better task understanding. These techniques are also employed in other LLM reasoning scenarios. Integrating these components, we design a meta prompt P_{meta} to query the LLM for n_0 rich views, forming an initial set of candidate views $\mathcal{V}^{(0)} = \{v_1^{(0)}, v_2^{(0)}, \ldots, v_{n_0}^{(0)}\}$, expressed as:

$$\mathcal{V}^{(0)} = \mathcal{M}(P_{\text{meta}}), \quad \mathcal{V}^{(0)} \in \mathbb{T}^{n_0} \tag{1}$$

where $\mathbb{T}$ denotes the set of textual segments.

To distill an effective view set, we select views through an iterative self-verification process. In this process, the LLM reassesses the initial view set $\mathcal{V}^{(0)}$ from an evaluator's perspective, verifying its relevance and filtering views that are genuinely beneficial for visual class identification. The verification process iteratively refines the view set $\mathcal{V}^{(t)}$ using the verification prompt P_{verify}. At each iteration t, an updated set $\mathcal{V}^{(t+1)}$ is produced, and the process stops when the set size stabilizes, indicating no further reduction:

$$\mathcal{V}^{(s+1)} = \mathcal{M}(P_{\text{verify}}(\mathcal{V}^{(s)})), \quad |\mathcal{V}^{(S)}| = |\mathcal{V}^{(S-1)}| \tag{2}$$

where $s = 0, 1, \ldots, S-1$, and S denotes the final iteration s. The converged view set is denoted as $\mathcal{V} = \mathcal{V}^{(S)}$, containing n views. This iterative self-verification ensures a robust view set, forming a reliable foundation for downstream tasks.

View-based Batch Querying Strategy

Based on the class set C and the view set $\mathcal{V}$, we can construct the class-view description matrix $S \in \mathbb{T}^{m \times n}$, where m is the size of C, and each element in S corresponds to a textual description. In terms of querying the LLM, a simple point-to-point query involves querying each class-view pair individually to populate the matrix S, and class-based querying focuses on querying one class across all views in a single round. However, due to the inherent randomness and quasi-independence of LLM outputs, both strategies can introduce biases across different classes under the same view, undermining inter-class comparability. Since our downstream tasks prioritize class-level comparability, we adopt the View-based Batch Querying Strategy. This approach retrieves descriptions for all classes under a single view in each round, preserving inter-class comparability across views.

We design a prompt P_{main} specifically tailored for view-based querying. The process can be expressed as follows:

$$S_j = \mathcal{M}(P_{\text{main}}(C, v_i)), \quad \forall j \in [1, n] \tag{3}$$

Finally, S is encoded by a pre-trained embedding model:

$$H = \mathcal{P}(S), \quad H \in \mathbb{R}^{m \times n \times d_t} \tag{4}$$

where d_t is the embedding dimension of the encoder. Open-source LLMs can also serve as embedding models, and we will validate their effectiveness in the experiments.

Inspired by [Liu *et al.*, 2024], we construct a global summary for each class using a global prompt P_{global} to extract comprehensive feature information. These features are embedded similarly to the multi-view descriptions, expressed as:

$$G = \mathcal{P}(\mathcal{M}(P_{\text{global}}(C))) \quad G \in \mathbb{R}^{m \times d_t} \tag{5}$$

where G represents the global representation of the classes. All parameters in both the LLM and embedding model remain frozen throughout the training process.

3.3 Description Refinement with Contextual Filtering Attention

We seek to extract task-relevant, effective class representations from the diverse and noisy descriptions derived from global and multiple local views. This is achieved through a contextual filtering attention module.

For each class c in C, we begin by projecting its global description G_c and multi-view local description H_c into the standard representations of the attention mechanism using linear transformations. Since G_c contains only global view, we expand its dimension into $\mathbb{R}^{n \times d_t}$ by copying the second dimension for n times in advance. Let $\mathcal{W}_Q, \mathcal{W}_K, \mathcal{W}_V$ be three MLPs that map the inputs to the query, key, and value representations, respectively. Specifically, the query, key, and value are defined as $Q = \mathcal{W}_Q(G_c)$, $K = \mathcal{W}_K(H_c)$, and $V = \mathcal{W}_V(H_c)$, respectively. Here, $Q, K, V \in \mathbb{R}^{n \times d_e}$, where d_e is the embedding dimension. In this design, the global description, containing high-level class information (e.g., key features), serves as the query, guiding the alignment of multi-view local descriptions.

The energy score $e \in \mathbb{R}^{n \times n}$ is computed based on a combination function $f(\cdot)$, which incorporates both the cosine similarity and the Euclidean distance between the query and key representations, formulated as follows:

$$e = f(Q, K) = \text{CosSim}(Q, K) \cdot ||Q - K||_2 \tag{6}$$

$$= \frac{\sum_{i=1}^{d_t} (Q^i \cdot K^i)}{|Q| \cdot |K|} \cdot \sqrt{\sum_{i=1}^{d_t} (Q^i - K^i)^2} \tag{7}$$

where $\text{CosSim}(\cdot)$ denotes cosine similarity and $||Q - K||_2$ represents the Euclidean distance between the Q and K. The cosine similarity and Euclidean distance complement each other, and their product simultaneously captures the global direction of the vectors and their local spatial differences, enabling better handling of complex relationships. The energy score reflects both the similarity and the spatial discrepancy between the query and key representations, leading to improved alignment of multi-view descriptions.

To transform the energy score e into a probability distribution, it is normalized into attention weights γ using a distribution function. In this case, we use the Softmax function to ensure that the weights sum to 1 and reflect the relative importance of each element in the sequence.

$$\gamma = g(e) = \text{Softmax}(e) = \frac{\exp(e)}{\sum_{j=1}^{n} \exp(e_j)} \tag{8}$$

Finally, we aggregate the value matrix V by performing matrix multiplication with the attention weights γ to obtain the weighted features:

$$T = \frac{1}{n} \sum_{i=1}^{n} (\gamma \cdot V)_i \tag{9}$$

The weighted feature T is a merged text-based embedding that captures the most discriminative features from the multi-view descriptions. Additionally, if an optional attribute A is available, it can be incorporated into the embedding representation T by concatenating it along the last dimension after dense embedding, thereby enhancing overall performance.

3.4 Class Injection with Cross-modal Co-learning

The fused text-based LLM-generated embedding, together with the vision-based classifier weight vector, is fed into a dual-autoencoder co-learning network. The network maps the text-based and vision-based information into a shared latent space $\mathcal{Z}$, enabling mutual supervision and integration of information from both sources. Specifically, the encoder for the vision-based embeddings is defined as $\mathcal{E}_v : \psi \rightarrow \mathcal{Z}$, with its corresponding decoder $\mathcal{D}_v : \mathcal{Z} \rightarrow \psi$. Similarly, the encoder for the text-based embeddings is $\mathcal{E}_t : T \rightarrow \mathcal{Z}$, with the decoder $\mathcal{D}_t : \mathcal{Z} \rightarrow T$. To enable the network to focus on the angular alignment of the injected weights, we adopt cosine distance as the distance metric $d(\cdot)$.

During the training stage, for each class, we strive to map both modalities of data into the same latent space by optimizing two types of loss: reconstruction loss within each embedding and cross-modal loss between both embeddings.

The reconstruction loss of the vision-based embedding aims to ensure the stability of the classifier weight embedding of seen classes, formulated as:

$$\mathcal{L}^s_{V\rightarrow V} = \sum_{c\in C_s} d(\mathcal{D}_v(\mathcal{E}_v(\psi^c)), \psi^c) \tag{10}$$

Similarly, the reconstruction loss for the text-based embedding, applied to both seen and unseen classes, is defined as:

$$\begin{aligned}\mathcal{L}_{T\rightarrow T} = &\sum_{c\in C}\sum_{v\in V} d(\mathcal{D}^v_t(\mathcal{E}_t(T^c)), L_v) \\ &+ \sum_{c\in C} d(\mathcal{D}^G_t(\mathcal{E}_t(T^c)), G)\end{aligned} \tag{11}$$

where an additional term $\sum_{c\in C} d(\mathcal{D}^A_t(\mathcal{E}_v(T^c)), A)$ can be added to $\mathcal{L}_{T\rightarrow T}$ if the class attribute is available.

The design of reconstruction losses ensures proper encoding in both text-based and vision-based latent spaces. However, enabling interactive learning between these two modalities requires alignment of their latent spaces. We achieve this goal through two cross-modal loss functions:

$$\mathcal{L}^s_{T\rightarrow V} = \sum_{c\in C_s} d(\mathcal{D}_v(\mathcal{E}_t(T^c)), \psi^c) \tag{12}$$

$$\begin{aligned}\mathcal{L}^s_{V\rightarrow T} = &\sum_{c\in C_s}\sum_{v\in V} d(\mathcal{D}^v_t(\mathcal{E}_v(\psi^c)), L^c_v) \\ &+ \sum_{c\in C_s} d(\mathcal{D}^G_t(\mathcal{E}_v(\psi^c)), G^c)\end{aligned} \tag{13}$$

where an additional term $\sum_{c\in C_s} d(\mathcal{D}^A_t(\mathcal{E}_v(\psi^c)), A^c)$ can be added to $\mathcal{L}^s_{V\rightarrow T}$ if the class attribute is available. The design of cross-modal loss facilitates latent space alignment in a bidirectional learning manner.

Finally, we add all reconstruction and cross-modal losses to obtain the total loss function $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}^s_{V\rightarrow V} + \mathcal{L}_{T\rightarrow T} + \mathcal{L}^s_{T\rightarrow V} + \mathcal{L}^s_{V\rightarrow T} \tag{14}$$

During the inference stage, we use the text-based embeddings to infer the expanded classifier weights $\hat{\psi}$, as follows:

$$\hat{\psi} = \hat{\mathcal{D}}_v(\hat{\mathcal{E}}_t(T)) \tag{15}$$

where $\hat{\mathcal{D}}_v$ is the trained vision-based decoder, and $\hat{\mathcal{E}}_t$ is the trained text-based encoder. The classifier expansion is completed once its weights are updated.

4 Experimental Study

Experiments on three common datasets are conducted. The results demonstrate the effectiveness and strong stability of CEMIL in both complementary and substitutive scenarios.

4.1 Benchmark Protocol

Datasets. Methods are evaluated on three widely used datasets: 1) *AWA2* [Xian *et al.*, 2018], an animal classification dataset featuring 50 mammal species; 2) *CUB* [Wah *et al.*, 2011], a dataset containing 200 bird species; and 3) *SUN* [Patterson *et al.*, 2014], a scene recognition dataset with 717 categories. Each dataset is accompanied by expert-constructed class attributes and is divided into seen and unseen classes based on the splitting scheme in [Xian *et al.*, 2017]. Following the image-free setting, we use only the part of class attributes here, without any images from these datasets.

Implementation Details. For each dataset, we pre-train a ResNet-101 [He *et al.*, 2016] using images and labels from only the seen classes, following [Xian *et al.*, 2019], as the base classifier for expansion. We utilize GPT-4o [OpenAI, 2023] as the LLM and the text encoder of CLIP [Radford *et al.*, 2021] as the embedding model, while subsequent analysis will demonstrate that this configuration has minimal impact on overall performance. The initial expected view number is set to 50. Each encoder is a single-layer MLP, while each decoder is a two-layer MLP with a hidden dimension 4096. The dimensions of the attention vectors are 2048. Neural network parameters are initialized randomly from a standard normal distribution. The Adam optimizer is used for training, with up to 500 epochs and an early stopping strategy. The learning rate is set to 1e-5, with batch sizes configured as 10, 16, and 32 for AWA2, CUB, and SUN, respectively. Experiments are conducted on an Nvidia GeForce RTX 4090 24GB GPU.

Evaluation Metrics. We perform evaluations of the methods under both standard and generalized ZSL settings. For standard ZSL, Top-1 accuracy is used as the metric, denoted as **T**. In generalized ZSL (GZSL), we calculate the accuracy for both unseen and seen classes, denoted as **u** and **s**, and compute their harmonic mean **H** by $2 \times (\mathrm{s} \times \mathrm{u})/(\mathrm{s} + \mathrm{u})$.

4.2 Comparison with State-of-the-Arts

The proposed CEMIL is compared with six image-free ZSL baselines: 1) *MLP* utilizes a two-layer neural network to map the class attributes to their corresponding weight vectors. 2) *ConSE* [Norouzi *et al.*, 2014] generates representation for unseen samples by computing a weighted sum of seen class embeddings, with the unseen sample's predicted probabilities as weights. 3) *COSTA* [Mensink *et al.*, 2014] also employs a weighted sum approach but uses co-occurrence statistics between classes as weights. 4) *SubReg* [Akyürek *et al.*, 2022] applies subspace regularization to keep unseen class weight vectors close to the subspace of seen classes, reducing catastrophic forgetting in CIL. 5) *VGSE* [Xu *et al.*, 2022] clusters local regions of seen classes by visual similarity and links these clusters to unseen classes via optimizing a class attributes similarity matrix. 6) *ICIS* [Christensen *et al.*, 2023] uses two separate encoder-decoders to predict unseen weights: one for class attributes and one for weight vectors.

Scenario	Method	AWA2				CUB				SUN			
		T	u	s	H	T	u	s	H	T	u	s	H
Expert-constructed Attribute only	MLP	46.8	2.0	95.9	4.0	41.4	0.0	87.6	0.0	49.7	0.0	50.1	0.0
	ConSE	44.0	3.0	96.1	5.7	41.9	0.5	88.0	0.9	44.4	0.1	47.9	0.1
	COSTA	40.9	0.0	96.1	0.0	31.9	0.0	87.6	0.0	19.9	0.0	50.1	0.0
	SubReg	37.5	0.0	96.1	0.0	37.6	0.0	87.6	0.0	48.3	0.0	50.1	0.0
	VGSE	55.4	31.8	92.4	47.3	45.1	39.2	52.3	44.8	42.7	42.5	1.6	3.1
	ICIS	64.6	35.6	93.3	51.6	60.6	45.8	73.7	56.5	51.8	45.2	25.6	32.7
LLM-based Supervision only	MLP	42.9	0.0	96.3	0.0	32.7	0.0	87.7	0.0	42.2	0.0	52.3	0.0
	ConSE	51.9	1.4	96.1	2.7	35.6	0.4	87.7	0.8	32.2	0.1	47.9	0.3
	COSTA	45.4	0.0	96.3	0.0	26.9	0.0	87.7	0.0	27.4	0.0	52.3	0.0
	SubReg	38.9	0.0	**96.3**	0.0	11.2	0.0	**87.7**	0.0	6.7	0.0	**52.3**	0.0
	VGSE	59.9	30.8	91.4	46.0	33.0	28.1	59.4	38.2	32.6	30.1	12.3	17.4
	ICIS	62.8	37.8	92.3	53.6	43.9	34.3	70.5	46.1	47.3	39.9	26.6	31.9
	CEMIL (Ours)	**69.1**	**41.4**	92.7	**57.3**	**61.8**	**44.6**	74.2	**55.7**	**55.1**	**46.5**	27.1	**34.2**
	Improvement	+10.0%	-	-	+6.9%	+41.7%	-	-	+20.8%	+16.4%	-	-	+7.2%
Expert-constructed Attribute + LLM	MLP	57.6	0.0	96.3	0.0	62.5	0.9	87.7	1.8	55.8	0.0	52.3	0.0
	ConSE	50.3	1.7	96.1	3.4	47.1	0.5	87.7	0.9	40.1	0.2	49.8	0.4
	COSTA	55.1	0.0	96.3	0.0	39.9	0.0	87.7	0.0	31.3	0.0	52.3	0.0
	SubReg	52.1	0.0	**96.3**	0.0	58.3	0.6	**87.7**	1.2	6.6	0.0	**52.3**	0.0
	VGSE	55.8	29.9	91.5	45.1	48.8	42.0	57.2	48.4	44.2	41.0	11.6	18.1
	ICIS	63.3	37.2	92.8	53.1	54.6	39.2	74.0	51.3	59.2	25.3	48.0	33.2
	CEMIL (Ours)	**73.2**	**48.6**	87.0	**62.4**	**68.3**	**53.2**	72.3	**61.3**	**60.7**	**40.7**	39.7	**40.2**
	Improvement	+15.6%	-	-	+17.5%	+25.1%	-	-	+19.5%	+2.5%	-	-	+21.1%

Table 1: Comparison of the proposed CEMIL with baseline methods. "T" indicates top-1 accuracy (%) in standard ZSL setting. In the generalized ZSL setting, "u" and "s" denote per-class accuracy (%) for unseen and seen test sets, respectively, and "H" is their harmonic mean. The best results are highlighted in **bold**. The "Improvement" row shows the percentage by which CEMIL surpasses SOTA methods.

The performance of these methods is evaluated across three scenarios: 1) *Expert-constructed Attribute only*: Utilizes attributes defined and annotated by domain experts, as provided in the ZSL datasets. 2) *LLM-based Supervision only*: Replaces expert-constructed attributes with multi-view descriptions generated using the proposed LLM-driven Robust Feature Enhancement method. For baseline methods, we ensure fairness by using the same multi-view embeddings as in CEMIL and applying mean pooling along the view to match the input shape of the attributes. 3) *Expert-constructed Attribute + LLM*: Combines both expert-constructed attributes and LLM-based supervision through concatenation.

Table 1 provides a comprehensive comparison of standard and generalized ZSL metrics across various methods. As demonstrated, CEMIL consistently outperforms state-of-the-art baselines across both scenarios and all three benchmarks, achieving significant improvements (mostly over 10%) in both standard and generalized ZSL settings. The performance of CEMIL, using both attribute and LLM supervision, reaches optimal levels, showcasing its effectiveness to complement existing expert-constructed attribute-based methods.

Notably, the proposed CEMIL, when using a more flexible LLM-based supervision, can achieve and even surpass SOTA methods that rely on expert-constructed attributes (*i.e.*, for ZSL: +6.9% on AWA2, +1.9% on CUB, +6.4% on SUN; for GZSL: +11.0% on AWA2, +4.6% on SUN). This highlights its ability to achieve superior performance with cost-effective LLMs, replacing traditional attribute-based methods, and underscores its potential to shift ZSL tasks from expert-dependent to LLM-driven approaches.

4.3 Ablation Study

To evaluate the contribution of key components in the proposed CEMIL framework, we perform a series of ablation experiments. The experiments are conducted by progressively modifying the full framework through the following steps: 1) Replacing hierarchical view reasoning in the DTG module with a plain prompt, and removing the process of iterative view selection. 2) Substituting the view-based text querying in the DTG module with a class-based querying approach. 3) Replacing the CFA module with a simplified version that integrates multi-view descriptions via mean pooling, concatenated with class attributes. 4) Replacing the CICC module with a single encoder-decoder architecture. 5) Removing the LLM supervision and relying solely on class attributes. As shown in Table 2, removing any module results in a performance decline, highlighting the importance and effectiveness of each proposed component throughout the workflow.

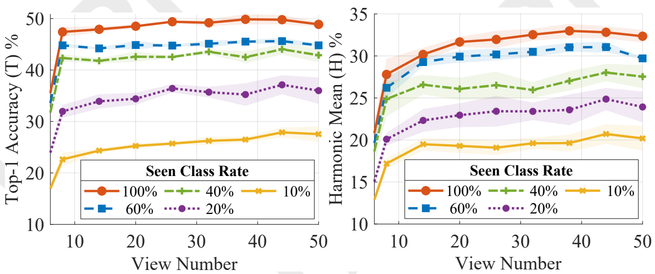
4.4 Empirical Analysis

Effect of the Number of Views and Seen Classes

In this experiment, we investigate the impact of the number of views and seen classes on model performance, as illustrated in Figure 3. The maximum number of views generated by the LLM is set to 50, with the number of views gradually increasing from 0. Additionally, we experiment with varying proportions of seen classes used for training, ranging from 20% to 100%. As the number of views increases, all performance metrics improve, and the rate of improvement slows beyond 20 views. Similarly, a larger number of seen classes has a positive impact on the final performance.

Setting	AWA2				CUB				SUN			
	T	u	s	H	T	u	s	H	T	u	s	H
CEMIL (Ours)	**73.2**	48.6	87.0	**62.4**	**68.3**	53.2	72.3	**61.3**	**60.7**	40.7	39.7	**40.2**
w/o DTG-hierarchical view reasoning	69.1	41.4	92.7	57.3	66.2	50.6	72.2	59.5	56.7	39.3	38.0	38.6
w/o DTG-view-based text querying	62.0	36.8	92.4	52.7	53.2	41.3	69.9	51.9	55.4	44.5	29.6	35.5
w/o CFA module	59.7	34.2	92.1	49.8	49.8	37.6	69.6	48.9	49.2	40.2	27.3	32.5
w/o CICC module	60.1	28.1	93.9	43.2	47.5	34.8	72.2	47.0	49.3	19.7	44.7	27.4
w/o LLM supervision	51.5	24.2	94.6	38.6	41.2	30.6	70.2	42.7	28.9	17.7	30.7	22.5

Table 2: Ablation study of the individual modules in CEMIL. We systematically remove each module—DTG (including hierarchical reasoning and view-based querying), CFA, CICC, and LLM supervision—and evaluate the performance under each configuration.



(a) ZSL Top-1 Accuracy (T) (b) GZSL Harmonic Mean (H)

Figure 3: Inference by the number of views and seen class rates in the SUN dataset. Experiments are conducted using LLM-based supervision only, without applying view selection.

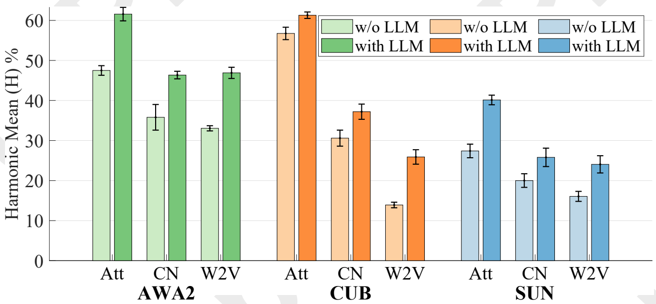


Figure 4: This figure illustrates the improvement in H achieved by combining LLMs with the proposed framework, highlighting the enhancements of CEMIL across various base knowledge sources.

Ability to Enhance Existing Knowledge Sources

To demonstrate the effectiveness of CEMIL in improving the efficiency of expert- or semi-expert-based sources, we conduct comparative experiments on different sources and datasets. We evaluate three knowledge sources: expert-constructed class attributes (Att), ConceptNet features [Speer and Lowry-Duda, 2017] (CN), and Wikipedia features [Yamada *et al.*, 2020] (W2V). Each experiment is conducted both with and without the assistance of the LLM, and the results are presented in Figure 4. Across all datasets and knowledge sources, CEMIL significantly improves the harmonic mean of zero-shot classifier expansion, highlighting its stable ability to enhance existing supervision sources.

Applicability of Different LLMs and Embedding Models

To validate the applicability of CEMIL across different LLMs, we replaced both the LLM and embedding components with different models. For the LLM, we evaluated four models, encompassing both open-source and proprietary options: GPT-4o [OpenAI, 2023], GPT-4o mini, LLaMA-3.1 [Touvron *et al.*, 2023], and Qwen-plus [Team, 2023]. For the embedding component, we explore the text encoder of CLIP [Radford *et al.*, 2021], SBERT [Reimers and Gurevych, 2019], and two open-source LLM-based embedding methods: LLaMA-3.1-8b and Qwen-2.5-7b. The LLM-based embedding is implemented by applying mean pooling to the vector from the last layer of the LLM, given the input text.

Embed Model	CLIP		SBERT		LLaMA		Qwen	
LLM	T	H	T	H	T	H	T	H
GPT-4o	60.7	40.2	58.6	39.4	58.5	38.1	54.5	37.5
GPT-4o mini	57.0	39.1	58.5	38.7	58.1	38.8	51.7	36.4
LLaMA-3.1	56.2	38.6	58.6	37.6	59.7	39.7	55.7	37.9
Qwen-plus	56.8	38.4	55.8	37.8	58.9	38.3	54.7	37.8

Table 3: Performance of the CEMIL method on the SUN dataset across different LLM and embedding model configurations. The CEMIL consistently achieves reliable results across all settings.

As shown in Table 3, all tested configurations of CEMIL perform well and consistently outperform previous attribute-based approaches. This demonstrates the robustness and broad applicability of CEMIL across various embedding models and LLM architectures. The configuration using LLaMA for both the LLM and the embedding model represents a scenario where classifier expansion is achieved solely through an offline open-source LLM, offering a highly flexible and practical choice for real-world applications.

5 Conclusion

This paper proposes CEMIL, a novel framework for expanding existing classifiers without the need for any images or human effort. CEMIL uses a structural two-stage strategy to deliver robust LLM supervision, integrates multi-view descriptions with contextual filtering attention, and employs a cross-modal co-learning framework to expand the classifier. Experiments show that CEMIL achieves state-of-the-art performance across multiple ZSL benchmarks in both standard and generalized image-free ZSL settings. The entire framework can be completed with only a pre-trained LLM, which can reshape the existing attribute-based ZSL paradigm in both complementary and substitutive scenarios. This work paves the way for flexible, cost-effective model adaptation to newly emerging classes in a fully automated manner.

Acknowledgments

This research was supported in part by the National Key R&D Program of China (No. 2021ZD0111700), in part by the National Nature Science Foundation of China (No. 62406302, 62137002, 62176245, 62206261), in part by the Natural Science Foundation of Anhui province (No. 2408085QF195), in part by the Fundamental Research Funds for the Central Universities under Grant (No. WK2150110035), in part by Special Foundation for Science and Technology Innovation and Entrepreneurship of CCTEG (No. 2020-2-TD-CXY006).

References

[Akyürek *et al.*, 2022] Afra Feyza Akyürek, Ekin Akyürek, Derry Wijaya, and Jacob Andreas. Subspace regularizers for few-shot class incremental learning. In *Proceedings of the International Conference on Learning Representations*, 2022.

[Ban *et al.*, 2025] Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, and Huanhuan Chen. LLM-driven causal discovery via harmonized prior. *IEEE Transactions on Knowledge and Data Engineering*, 37(4):1943–1960, 2025.

[Christensen *et al.*, 2023] Anders Christensen, Massimiliano Mancini, A Koepke, Ole Winther, and Zeynep Akata. Image-free classifier injection for zero-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19072–19081, 2023.

[Guo *et al.*, 2023] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 10867–10877, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 770–778, 2016.

[Hersche *et al.*, 2022] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 9057–9067, 2022.

[Lampert *et al.*, 2013] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2013.

[Li *et al.*, 2023] Xiaofan Li, Yachao Zhang, Shiran Bian, Yanyun Qu, Yuan Xie, Zhongchao Shi, and Jianping Fan. VS-Boost: Boosting visual-semantic association for generalized zero-shot learning. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pages 1107–1115, 2023.

[Li *et al.*, 2024] Qian Li, Zhuo Chen, Cheng Ji, Shiqi Jiang, and Jianxin Li. LLM-based multi-level knowledge generation for few-shot knowledge graph completion. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 2135–2143, 2024.

[Liu *et al.*, 2024] Yicheng Liu, Jie Wen, Chengliang Liu, Xiaozhao Fang, Zuoyong Li, Yong Xu, and Zheng Zhang. Language-driven cross-modal classifier for zero-shot multi-label image recognition. In *Proceedings of the International Conference on Machine Learning*, volume 235, pages 32173–32183. PMLR, 2024.

[Mensink *et al.*, 2014] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. COSTA: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 2441–2448, 2014.

[Mittal *et al.*, 2021] Sudhanshu Mittal, Silvio Galesso, and Thomas Brox. Essentials for class incremental learning. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 3513–3522, 2021.

[Naeem *et al.*, 2023] Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2MVFormer: Large language model generated multi-view document supervision for zero-shot image classification. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 15169–15179, 2023.

[Naeem *et al.*, 2024] Muhammad Ferjad Naeem, Yongqin Xian, Luc Van Gool, and Federico Tombari. I2DFormer+: Learning image to document summary attention for zero-shot image classification. *International Journal of Computer Vision*, 132(9):3806–3822, 2024.

[Norouzi *et al.*, 2014] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *Proceedings of the International Conference on Learning Representations*, 2014.

[OpenAI, 2023] OpenAI. GPT-4. https://openai.com/index/gpt-4-research/, 2023. Accessed: 2023-03-14.

[Patterson *et al.*, 2014] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The SUN attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108:59–81, 2014.

[Qiao *et al.*, 2016] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton Van Den Hengel. Less is more: Zero-shot learning from online textual documents with noise suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2249–2257, 2016.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack

Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139, pages 8748–8763. PMLR, 2021.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics, 2019.

[Saha *et al.*, 2024] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting VLMs with text descriptions. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 17542–17552, 2024.

[Speer and Lowry-Duda, 2017] Robyn Speer and Joanna Lowry-Duda. ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, 2017.

[Team, 2023] Qwen Team. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. *Computation & Neural Systems Technical Report*, 2011.

[Wang *et al.*, 2024] Shuai Wang, Yibing Zhan, Yong Luo, Han Hu, Wei Yu, Yonggang Wen, and Dacheng Tao. Joint input and output coordination for class-incremental learning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 5108–5116, 2024.

[Wei *et al.*, 2021] Kun Wei, Cheng Deng, Xu Yang, and Dacheng Tao. Incremental zero-shot learning. *IEEE Transactions on Cybernetics*, 52(12):13788–13799, 2021.

[Wu *et al.*, 2024] Xingyu Wu, Yan Zhong, Jibin Wu, Bingbing Jiang, and Kay Chen Tan. Large language model-enhanced algorithm selection: Towards comprehensive algorithm representation. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 5235–5244, 2024.

[Xian *et al.*, 2017] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017.

[Xian *et al.*, 2018] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2018.

[Xian *et al.*, 2019] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. F-VAEGAN-D2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition*, pages 10275–10284, 2019.

[Xu *et al.*, 2020] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33:21969–21980, 2020.

[Xu *et al.*, 2022] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. VGSE: Visually-grounded semantic embeddings for zero-shot learning. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 9316–9325, 2022.

[Yamada *et al.*, 2020] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 23–30, 2020.

[Yun *et al.*, 2023] Sukmin Yun, Seong Hyeon Park, Paul Hongsuck Seo, and Jinwoo Shin. IFSeg: Image-free semantic segmentation via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2967–2977, 2023.

[Zhao *et al.*, 2024] Zengqun Zhao, Yu Cao, Shaogang Gong, and Ioannis Patras. Enhancing zero-shot facial expression recognition by LLM knowledge transfer. *arXiv preprint arXiv:2405.19100*, 2024.

[Zhou *et al.*, 2024] Haichen Zhou, Yixiong Zou, Ruixuan Li, Yuhua Li, and Kui Xiao. Delve into base-novel confusion: Redundancy exploration for few-shot class-incremental learning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 5635–5643, 2024.

[Zhu *et al.*, 2018] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1004–1013, 2018.