Hierarchical Semantic Alignment for Image Clustering

Xingyu Zhu[1, 2], **Beier Zhu**[2], **Yunfan Li**[3], **Junfeng Fang**[4], **Shuo Wang**[1*], **Kesen Zhao**[2], **Hanwang Zhang**[2]

[1] University of Science and Technology of China
[2] Nanyang Technological University
[3] Sichuan University
[4] National University of Singapore
xingyuzhu@mail.ustc.edu.cn, shuowang.edu@gmail.com

Abstract

Image clustering is a classic problem in computer vision, which categorizes images into different groups. Recent studies utilize nouns as external semantic knowledge to improve clustering performance. However, these methods often overlook the inherent ambiguity of nouns, which can distort semantic representations and degrade clustering quality. To address this issue, we propose a hierar**C**hical sem**A**ntic alignm**E**nt method for image clustering, dubbed **CAE**, which improves clustering performance in a training-free manner. In our approach, we incorporate two complementary types of textual semantics: caption-level descriptions, which convey fine-grained attributes of image content, and noun-level concepts, which represent high-level object categories. We first select relevant nouns from WordNet and descriptions from caption datasets to construct a semantic space aligned with image features. Then, we align image features with selected nouns and captions via optimal transport to obtain a more discriminative semantic space. Finally, we combine the enhanced semantic and image features to perform clustering. Extensive experiments across 8 datasets demonstrate the effectiveness of our method, notably surpassing the state-of-the-art training-free approach with a 4.2% improvement in accuracy and a 2.9% improvement in adjusted rand index (ARI) on the ImageNet-1K dataset.

1 Introduction

Image clustering (Likas, Vlassis, and Verbeek 2003) is a foundational task in computer vision that aims at grouping images into clusters, where instances from the same cluster share similar semantics. In the era of deep learning, a series of works center on learning discriminative features (Qiu et al. 2025; Zhu et al. 2025a) for clustering, such as contrastive learning (Li et al. 2021, 2022) and self-supervised learning (He et al. 2020; Chen et al. 2020). The recent progress of deep clustering has reached a plateau, as the internal priors from the images themselves provide limited improvement. To address this, recent methods seek to incorporate external priors, such as text, to enhance clustering performance. For instance, Cai *et al.* (Cai et al. 2023) map images into a semantic space constructed from related nouns in WordNet (Miller 1995). Similarly, Li *et al.* (Li et al. 2024) combine noun embeddings with image embeddings to facilitate clustering.

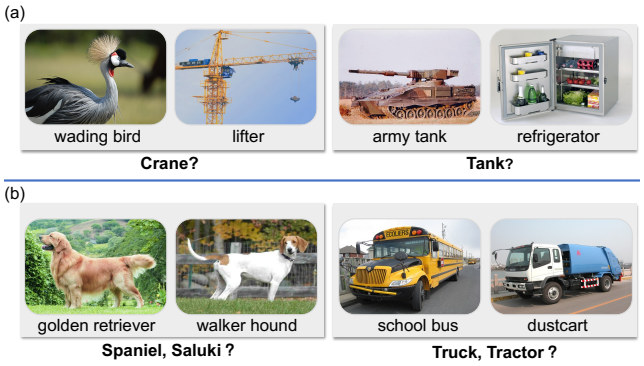


Figure 1: Observations of ambiguity in a single noun from the ImageNet (Deng et al. 2009) dataset. In (a), the words "crane" and "tank" can refer to entirely different objects, respectively. In (b), the semantic similar words "spaniel", "saluki", "truck", and "tractor" fail to distinguish the fine-grained classes.

Although these methods leverage the rich semantics encoded in textual data, addressing that the inclusion of external knowledge can enhance the ability to differentiate between visually similar clusters, they overlook the inherent ambiguities of nouns. These ambiguities can ultimately diminish the effectiveness of clustering, particularly in cases where the same noun can refer to multiple concepts or subclasses. As shown in Figure 1 (a), the noun "crane" can refer to a bird or a lifting machine, and the "tank" could denote a military vehicle or a refrigerator. Additionally, Figure 1 (b) illustrates that a similar noun fails to represent the specific characteristics of fine-grained classes. For example, the term "spaniel" or "saluki" does not distinguish between a "golden retriever" and a "walker hound," nor does the noun "truck" or "tractor" clearly differentiate between a "school bus" and a "dustcart." Therefore, relying solely on nouns as external knowledge for clustering struggles to provide precise semantics, especially when faced with ambiguities arising from polysemy or multiple subclasses.

To address these challenges, we propose a training-free method, dubbed **CAE**, that combines caption-based semantics from image descriptions (Plummer et al. 2015; Yang et al. 2023; Wu et al. 2025) with noun-based semantics from WordNet (Miller 1995) for guided image clustering. These two

*Corresponding author

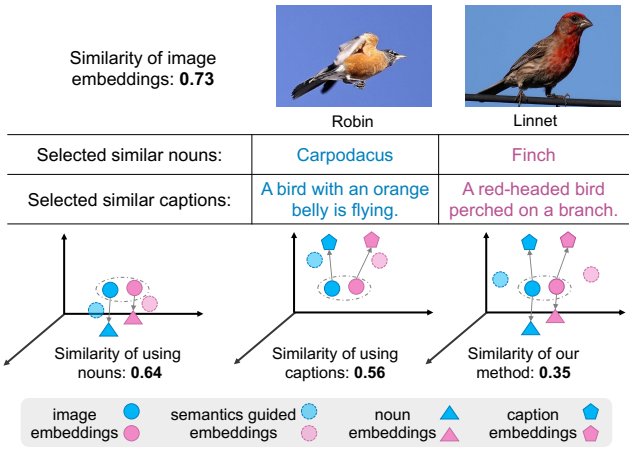


Figure 2: Comparison of embeddings similarity using nouns, captions, and our proposed method for two similar bird images ("Robin" and "Linnet"). Despite the high similarity in image embeddings (0.73), using only nouns or captions yields higher semantic similarity (0.64 and 0.56, respectively). By combining both nouns and captions, our method reduces the similarity score to 0.35, providing a more accurate distinction between the images.

types of semantics are complementary: nouns capture high-level object categories, while captions provide fine-grained attribute details of the image content. By combining these two forms of semantics, we can achieve a more comprehensive understanding of the image. As illustrated in Figure 2, we compare the effectiveness of using only nouns, only captions, and our proposed combination of both. Consider two visually similar bird images from different classes: a robin and a linnet with a cosine similarity of 0.73 between their image embeddings (solid circle). When using only nouns, we select the most similar terms from WordNet: "Carpodacus" for the robin and "Finch" for the linnet. These nouns (triangle) are relevant birds but are not the true labels, so they introduce some discrepancy and result in decreasing the similarity from 0.73 to 0.64, as the categorical distinctions between the nouns capture only part of the semantic difference between the images. When using captions alone, "A bird with an orange belly is flying" for the robin and "A red-headed bird perched on a branch" for the linnet. These captions (pentagon) emphasize the differences in the birds' characteristics (orange belly versus red-headed) but retain the common object bird, contributing to reducing the similarity to 0.56. Although captions offer a contextual distinction, the shared object limits their ability to differentiate the images fully. However, our method combines both nouns and captions, achieves a more distinguishable representation by leveraging both the categorical distinction from nouns and the contextual details from captions, and leads to a significant reduction in similarity, from 0.73 to 0.35.

Specifically, our method consists of two components: semantic space construction and semantic space interaction. In the semantic space construction stage, we first assign all nouns and captions to image semantic centers and then select relevant nouns and captions to construct the counterparts for images. In the adaptive semantics fusion module, we enable image, noun, and caption features to interact through prototype-guided weighting based on their semantic similarity. Finally, the fused representations are used to guide clustering in a more discriminative way.

The main contributions are summarized as follows:

- We propose a training-free method that exploits the external semantic knowledge from both nouns and captions to effectively guide the clustering process.
- We first construct a semantic space for images by aligning its distribution with relevant nouns and descriptions. Then we design an adaptive fusion strategy by leveraging this space to represent the images.
- We demonstrate the effectiveness of our proposed method through extensive experiments on five classic datasets and three challenging datasets, showing consistent and significant improvements over existing baselines, including outperforming zero-shot CLIP.

2 Related Works

In this section, we begin with a concise review of deep clustering methods, followed by a discussion on pre-trained vision-language models and their applications.

2.1 Deep Clustering

To tackle high-dimensional real-world data, deep clustering methods leverage neural networks to learn discriminative features, enabling the effective categorization of samples into distinct clusters (Peng et al. 2016; Yang, Parikh, and Batra 2016; Xie, Girshick, and Farhadi 2016; Peng et al. 2018). Early deep clustering research focuses on adapting classic clustering objectives (Yang et al. 2017; Ji et al. 2017; Shaham and Stanton 2018) into loss functions to optimize neural networks. Recently, motivated by the success of contrastive learning (Chen et al. 2020; He et al. 2020), contrastive clustering methods leverage the augmentation invariance at instance- (Hu et al. 2017; Ji, Henriques, and Vedaldi 2019) or cluster-level (Li et al. 2021; Huang, Gong, and Zhu 2020) to achieve end-to-end clustering. Another branch of research resorts to pseudo-labeling techniques to further boost the clustering performance (Van et al. 2020; Li et al. 2022; Niu, Shan, and Wang 2022). Notably, different from previous multi-modal clustering methods (Gao et al. 2020; Mao et al. 2021) that require paired image-text data as input, externally-guided clustering method (Cai et al. 2023; Li et al. 2024) does not rely on prior knowledge such as the pairing information and aims at exploring broader external knowledge.

Our work further explores the externally-guided clustering paradigm, advancing both the scope and application of external knowledge. Compared to existing works that leverage semantic information through cross-modal distillation (Li et al. 2024) or pseudo-labeling (Cai et al. 2023), we demonstrate that image caption models offer promising guidance for clustering. Moreover, We propose a more effective strategy to incorporate image and textual semantics, through the optimal transport and a residual attention mechanism.

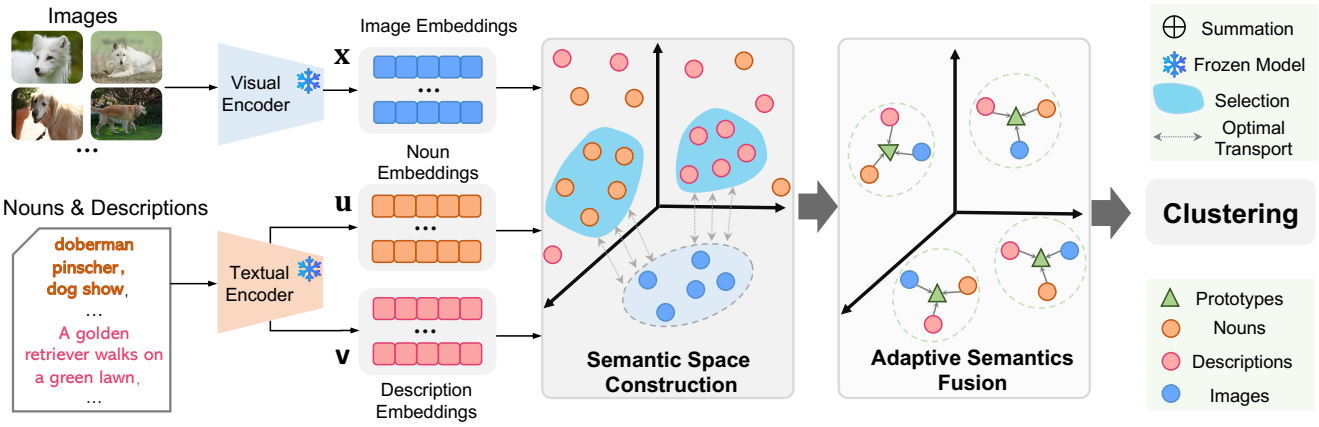


Figure 3: An overview of our method, which consists of two components, *i.e.*, (a) **Semantic Space Construction:** Select the nouns and descriptions that include the same semantics with image embeddings. (b) **Adaptive Semantics Fusion:** The selected nouns and descriptions embeddings are leveraged to boost image semantics through adaptive fusion.

2.2 Zero/Few-shot Classification

With advances in pre-training techniques, vision-language pre-training models (VLM), such as CLIP (Radford et al. 2021), have demonstrated a remarkable ability to align images and texts within a unified feature space. Thanks to the incorporation of textural semantics, these models readily adapt to a variety of downstream tasks, such as classification (Martin et al. 2024; Zhu et al. 2023, 2025b), in few-shot (Tang et al. 2024; Zhu et al. 2024c) or even zero-shot (Novack et al. 2023; Ge et al. 2023; Conti et al. 2023; Menon and Vondrick 2023; Zhu et al. 2024b) scenarios. For example, they perform classification by calculating the similarity between the test image and predefined class names. To better leverage the semantics in class names, Menon *et al.* (Menon and Vondrick 2023) proposed prompting a large language model to generate distinct descriptions for specific classes. Novack *et al.* (Novack et al. 2023) enriched the class names through a hierarchical fine-grained label set. It is worth noting that Conti *et al.* (Conti et al. 2023) addresses the open vocabulary classification problem by using a multimodal large language model (MVLM) to generate descriptions and then parse class names for each image. However, most of these methods rely heavily on the availability of class names as prior knowledge, which may not always be accessible in real-world scenarios, particularly in clustering tasks. As a result, the potential of the text modality in VLMs cannot be fully utilized, leading to sub-optimal performance.

In this work, we present a more flexible approach to using VLMs by incorporating nouns and descriptions that are not predefined, offering a solution to the challenge of lacking class name prior knowledge. Specifically, we propose an advanced text space construction strategy that effectively captures and leverages visual semantics, thereby enhancing visual representations for image clustering tasks. We hope this strategy will inspire broader applications of VLMs in fully unsupervised tasks.

3 Method

In this section, we present our training-free method as illustrated in Figure 3. We begin by introducing the settings and notations used throughout our approach in Section 3.1. Following this, we describe our method in detail, including the construction of the semantic space in Section 3.2 and the semantics fusion in Section 3.3.

3.1 Preliminaries

Giving an image dataset $\{x_i\}_{i=1}^N$ with N samples, whose embeddings are computed as $\mathbf{x}_i = \Phi_{\mathrm{v}}(x_i)$, where Φ_{v} represents CLIP's visual encoder. To capture the semantics of these images, we introduce two textual datasets. The first is a nouns dataset $\{\mathbf{u}_i\}_{i=1}^T$ containing T words from WordNet (Miller 1995), a large English lexical database that groups words into sets of synonyms called synsets. The second is a captions dataset $\{\mathbf{v}_i\}_{i=1}^M$ with M captions from Flickr (Plummer et al. 2015; Fang et al. 2022), an online photo-sharing platform where images are often accompanied by user-generated captions. The noun embeddings $\mathbf{u}_i = \Phi_{\mathrm{t}}(u_i)$ and caption embeddings $\mathbf{v}_i = \Phi_{\mathrm{t}}(v_i)$ are calculated using CLIP's textual encoder Φ_{t}. The image embeddings and textual embeddings share the same dimension, where $\mathbf{x}_i, \mathbf{u}_i, \mathbf{v}_i \in \mathbb{R}^d$. The goal is to assign the images into K clusters.

3.2 Semantic Space Construction

Filtering nouns and descriptions. External semantic knowledge has been shown to enhance the performance of image classification and clustering tasks (Niu, Shan, and Wang 2022; Radford et al. 2021; Li et al. 2024; Cai et al. 2023; Zhu et al. 2024a). In this paper, we leverage textual semantics to assist clustering tasks. Notably, clustering tasks lack priors of object names or object attributes for individual images. To address this, we construct a semantic space by selecting a subset of nouns and captions from WordNet and Flickr that encompass the semantics of objects and attributes, enabling the description of images in the text modality.

A well-constructed semantic space should focus on image-relevant semantics while excluding unrelated ones to prevent confusion in clustering. A small space risks losing critical discriminative information, whereas a large one may introduce excessive general semantics, diminishing clustering effectiveness. To ensure an appropriate size, we follow the setting in (Li et al. 2024): the dataset is clustered into $n = N/300$ groups, and the closest nouns and captions are retrieved for each cluster to define the semantic space. Specifically, we apply k-means clustering on the image embeddings and compute the semantic center of the j-th cluster:

$$\mathbf{p}_j = \frac{1}{|\mathcal{P}_j|} \sum_{i \in \mathcal{P}_j} \mathbf{x}_i,\ j \in [1, n], \tag{1}$$

where $\mathcal{P}_j$ denotes the set of images assigned to the j-th cluster. To identify representative nouns, we first define the probability of the i-th noun belonging to the j-th cluster as:

$$p(\mathbf{p}_j|\mathbf{u}_i) = \frac{\exp(\mathbf{u}_i^\top \mathbf{p}_j)}{\sum_{l=1}^{n} \exp(\mathbf{u}_i^\top \mathbf{p}_l)}. \tag{2}$$

We reorder the nouns embeddings according to decreasing probability $p(\mathbf{p}_j|\mathbf{u}_i)$, denoting the reordered sequence as $[\mathbf{u}_{(1)}, ...\mathbf{u}_{(i)}, ..., \mathbf{u}_{(T)}]$. We then retrieve the top-K nouns candidates with the highest probability for the j-th cluster:

$$\mathcal{U}_j = \{\mathbf{u}_{(i)} |\ i \leq K\}. \tag{3}$$

Finally, we take the union of these sets to form the selected nouns across all clusters: $\mathcal{U} = \bigcup_{j=1}^{n} \mathcal{U}_j$. The closest captions, $\mathcal{V}$, for all clusters are in the same manner.

Constructing counterparts. The selected nouns and captions capture different aspects of image semantics: nouns typically represent general object names, and captions often include fine-grained attributes. Leveraging these selected embeddings, $\mathcal{U}$ and $\mathcal{V}$, we aim to compute the noun counterpart and the caption counterpart for each image $\mathbf{x}_i$ by assessing the distributional distances between image and textual embeddings. To achieve this, we employ Optimal Transport (OT) (Villani et al. 2009), to aligning their distributions. Mathematically, we define two empirical distributions P and Q to model the sets of two modalities:

$$P = \sum_{i=1}^{N} \frac{1}{N} \delta_{\mathbf{x}_i},\ Q = \sum_{j=1}^{|\mathcal{U}|} \frac{1}{|\mathcal{U}|} \delta_{\mathbf{u}_j}, \tag{4}$$

where $\delta_{\mathbf{x}_i}$ and $\delta_{\mathbf{u}_j}$ are the Dirac delta function centered at $\mathbf{x}_i$ and $\mathbf{u}_j$, respectively. The OT distance between P and Q is thus defined as:

$$d_{\mathrm{OT}}(P, Q; \mathbf{C}) := \min_{\mathbf{T} \in \Pi(P,Q)} \langle \mathbf{T}, \mathbf{C} \rangle, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius dot-product, $\mathbf{T} \in \mathbb{R}^{N \times |\mathcal{U}|}$ is the transport plan, and $\mathbf{C} \in \mathbb{R}^{N \times |\mathcal{U}|}$ the cost matrix. Each element $c_{i,j}$ is calculated using cosine similarity:

$$c_{i,j} = 1 - s^u_{i,j}, \tag{6}$$

where $s^u_{i,j} = \mathbf{x}_i \top \mathbf{u}_j / (||\mathbf{x}_i|| \cdot ||\mathbf{u}_j||)$. The goal of OT is to minimize the total cost to transporting mass from $\mathbf{x}$ to $\mathbf{u}$. To achieve this, we use the Sinkhorn-Knopp algorithm (Cuturi 2013) to approximate the solution $\mathbf{T}^u$. After obtaining the transport plan $\mathbf{T}^u$, we compute the noun counterpart $\mathbf{x}^u_i$ for each image $\mathbf{x}_i$:

$$\mathbf{x}^u_i = \sum_{j=1}^{|\mathcal{U}|} t^u_{i,j} s^u_{i,j} \mathbf{u}_j, \tag{7}$$

where $t_{i,j}$ is the i-th column and j-th row element in $\mathbf{T}^u$. Following the same procedure, we also compute the caption counterpart $\mathbf{x}^v_i$ for each image $\mathbf{x}_i$:

$$\mathbf{x}^v_i = \sum_{j=1}^{|\mathcal{V}|} t^v_{i,j} s^v_{i,j} \mathbf{v}_j, \tag{8}$$

where $t^v_{i,j}$ is the element of optimal transport plan $\mathbf{T}^v$ between $\mathbf{x}_i$ and $\mathbf{v}_j$. While optimal transport offers a principled matching strategy, an alternative and commonly used approach is to apply similarity-based softmax weighting. Specifically, one may compute $\mathbf{x}^u_i = \sum_{j=1}^{|\mathcal{U}|} w^u_{i,j} \mathbf{u}_j$, with softmax weights $w^u_{i,j} = \frac{\exp(s^u_{i,j})}{\sum_{k=1}^{|\mathcal{U}|} \exp(s^u_{i,k})}$. We compare this similarity+softmax approach to OT in the following theorem.

Theorem 1. *Let $\{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{R}^d$ be a set of image embeddings, and let $\{\mathbf{u}_j\}_{j=1}^{|\mathcal{U}|} \subset \mathbb{R}^d$ denote a set of textual embeddings (e.g., noun or caption representations). For clarity, we present the theoretical analysis using noun embeddings as a representative case. The softmax-based aggregation:*

$$\mathbf{x}^u_i = \sum_{j=1}^{|\mathcal{U}|} w^u_{i,j} \mathbf{u}_j, \quad w^u_{i,j} = \frac{\exp(s^u_{i,j})}{\sum_{k=1}^{|\mathcal{U}|} \exp(s^u_{i,k})}, \tag{9}$$

is a special case of entropic OT when the column-wise marginal constraint $\sum_i t^u_{i,j} = \frac{1}{|\mathcal{U}|}$ is relaxed. For any $\delta > 0$, with probability at least $1 - \delta$ over Sinkhorn iterations, softmax incurs higher semantic error than OT.

Proof. The entropic OT formulation defines transport weights as:

$$t^u_{i,j} = a_i b_j \exp\left(-\frac{c_{i,j}}{\epsilon}\right), \quad c_{i,j} = 1 - s^u_{i,j}, \tag{10}$$

where a_i, b_j are positive scaling factors. These factors are determined via Sinkhorn iterations such that the transport matrix $\mathbf{T}^u$ satisfies the marginal constraints:

$$\sum_j t^u_{i,j} = \frac{1}{N}, \quad \sum_i t^u_{i,j} = \frac{1}{|\mathcal{U}|}. \tag{11}$$

Softmax weights correspond to:

$$t^u_{i,j} = \frac{w^u_{i,j}}{N}, \quad w^u_{i,j} = \frac{\exp(s^u_{i,j})}{\sum_k \exp(s^u_{i,k})}, \tag{12}$$

which satisfies $\sum_j t^u_{i,j} = \frac{1}{N}$, but generally violates the column-wise constraint $\sum_i t^u_{i,j} = \frac{1}{|\mathcal{U}|}$, i.e., noun embeddings are not used in a balanced way. As for optimization, softmax minimizes a per-image cost $\sum_j w^u_{i,j} c_{i,j} + \sum_j w^u_{i,j} \log w^u_{i,j}$

subject to $\sum_j w^u_{i,j} = 1$, while OT globally minimizes a regularized cost under marginal constraints. Hence,

$$E_{\text{softmax}} = \sum_{i,j} \frac{w^u_{i,j}}{N} c_{i,j} \geq \sum_{i,j} t^u_{i,j} c_{i,j} = E_{\text{OT}}, \tag{13}$$

as OT minimizes Eq. (5). With probability at least $1 - \zeta$ over Sinkhorn iterations (Cuturi 2013), OT's column constraint, via b_j, balances noun usage, reducing intra-cluster variance $\mathbb{E}_{(i,i')\in Z_k}\left[\left\|\sum_j t^u_{i,j}\mathbf{u}_j - \sum_j t^u_{i',j}\mathbf{u}_j\right\|^2\right] \leq \mathbb{E}_{(i,i')\in C_k}\left[\left\|\sum_j w^u_{i,j}\mathbf{u}_j - \sum_j w^u_{i',j}\mathbf{u}_j\right\|^2\right]$ for cluster Z_k, enhancing clustering. □

3.3 Adaptive Semantics Fusion

Given multimodality features $\mathbf{x}_i$, $\mathbf{x}^u_i$, and $\mathbf{x}^v_i$ extracted from image, noun, and caption modalities respectively, a naive strategy is to concatenate them into $\{[\mathbf{x}_i; \mathbf{x}^u_i; \mathbf{x}^v_i]\}_{i=1}^N$ and perform k-means clustering. Although effective to some extent (Section 4.3), this approach neglects the varying contribution and interaction among modalities. To enable adaptive feature fusion, we introduce a prototype-guided weighting mechanism. Let $\mathbf{x}_i$, $\mathbf{x}^u_i$, and $\mathbf{x}^v_i$ represent the image, noun, and caption features for the i-th instance. We define a unified modality set $\mathcal{M} = \{(m) : m \in \{\mathbf{x}_i, \mathbf{x}^u_i, \mathbf{x}^v_i\}\}$ and compute the semantic prototype by averaging over all modalities:

$$\mathbf{x}^p_i = \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} \mathbf{x}^{(m)}_i. \tag{14}$$

Then, we compute the cosine similarity between each modality-specific feature and the prototype:

$$\alpha^{(m)}_i = \frac{\langle \mathbf{x}^{(m)}_i, \mathbf{x}^p_i \rangle}{\|\mathbf{x}^{(m)}_i\| \cdot \|\mathbf{x}^p_i\|}, \quad \forall m \in \mathcal{M}. \tag{15}$$

Let $\boldsymbol{\alpha}_i = [\alpha^{(m)}_i]_{m\in\mathcal{M}}$ be the similarity vector. We apply temperature-scaled softmax to obtain the modality weights:

$$\boldsymbol{\beta}_i = \text{softmax}\left(\boldsymbol{\alpha}_i/\gamma\right), \quad \text{with } \boldsymbol{\beta}_i = [\beta^{(m)}_i]_{m\in\mathcal{M}}. \tag{16}$$

Finally, the fused representation for each instance is obtained via a weighted combination of modality-specific features:

$$\bar{\mathbf{x}}_i = \sum_{m\in\mathcal{M}} \beta^{(m)}_i \mathbf{x}^{(m)}_i. \tag{17}$$

This formulation enables instance-wise adaptive fusion, allowing the model to emphasize modalities that are more semantically aligned with the visual content. Such content-aware integration facilitates fine-grained cross-modal representation learning. Subsequently, we perform clustering on the fused embeddings using the standard k-means algorithm:

$$y_i := \text{k-means}(\bar{\mathbf{x}}_i), \quad i \in [1, N], \tag{18}$$

where y_i denotes the predicted cluster assignment for the i-th image. The complete procedure of our methods is summarized in Algorithm 1.

Algorithm 1: Pipeline of **CAE**

Input: Image embeddings $\{\mathbf{x}_i\}_{i=1}^N$, noun embeddings $\{\mathbf{u}_i\}_{i=1}^T$, description embeddings $\{\mathbf{v}_i\}_{i=1}^M$.

Step1: construct semantic space

1: Construct relevant nouns set $\mathcal{U}$ and descriptions set $\mathcal{V}$ by Eq. (2) and Eq. (3).
2: Build transport plan $\mathbf{T}^u$ and $\mathbf{T}^v$ by optimize Eq. (5).
3: Compute counterparts $\{\mathbf{x}^u_i\}_{i=1}^N$ and $\{\mathbf{x}^v_i\}_{i=1}^N$ by Eq. (7) and Eq. (8), respectively.

Step2: adaptive semantics fusion

4: Compute similarity weights $\boldsymbol{\alpha}$ by Eq. (15).
5: Compute fused features $\bar{\mathbf{x}}_i$ by Eq. (17).

Output: Compute cluster assignments $\{y_i\}_{i=1}^N$ by Eq. (18).

4 Experiments

In this section, we present the experimental evaluation of our method, including performance comparisons, ablation studies, and clustering visualizations.

4.1 Setup

Datesets. We conduct experiments on five widely used datasets: STL10 (Coates, Ng, and Lee 2011), CIFAR-10 (Krizhevsky and Hinton 2009), CIFAR-20 (Krizhevsky and Hinton 2009), ImageNet-10 (Chang et al. 2017), and ImageNet-Dogs (Chang et al. 2017), and three challenging datasets: DTD (Cimpoi et al. 2014), UCF-101 (Soomro, Zamir, and Shah 2012), and ImageNet-1K (Deng et al. 2009). These datasets cover a diverse range of image categories, allowing us to assess the generalizability of clustering methods across diverse scenarios.

Evaluation Metrics. We utilize three popular metrics to evaluate the clustering performance, including Normalized Mutual Information (NMI), Accuracy (ACC), and Adjusted Rand Index (ARI). The higher values of these metrics indicate the better performance.

Implementation details. To ensure a fair comparison with prior work SIC (Cai et al. 2023) and TAC (Li et al. 2024), we use the CLIP model (Radford et al. 2021) with ViT-B/32 as the image encoder and a Transformer as the text encoder. We assemble the nouns from WordNet in the template as "a photo of [CLASS]," and the captions are filtered in the same manner as used in (Fang et al. 2022). The temperature parameter γ is set as 0.01, for all datasets. All experiments are conducted on a single Nvidia RTX 3090 GPU.

4.2 Main Results

We compare our approach with state-of-the-art (SOTA) methods and categorize the evaluated methods into training-based and training-free. We first evaluate our method on 5 classic clustering datasets, comparing it with 19 training-based methods and 2 training-free methods. Previous works have utilized various backbones, such as ResNet-34 (SPICE (Niu, Shan, and Wang 2022), TCL (Li et al. 2022)) and ResNet-18 (SPICE (Niu, Shan, and Wang 2022), SCAN (Van et al. 2020)). With the rapid development of model pre-training, CLIP has been adopted in clustering tasks (Li et al. 2024; Cai et al. 2023). The comparison includes a wide range of representative methods, such as PICA (Huang, Gong, and Zhu

	Method	STL-10			CIFAR-10			CIFAR-20			ImageNet-10			ImageNet-Dogs			AVG
		NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
Training-based	PICA (CVPR20)	61.1	71.3	53.1	59.1	69.6	51.2	31.0	33.7	17.1	80.2	87.0	76.1	35.2	35.3	20.1	52.1
	IDFD (ICLR20)	64.3	75.6	57.5	71.1	81.5	66.3	42.6	42.5	26.4	89.8	95.4	90.1	54.6	59.1	41.3	63.9
	SCAN (ECCV20)	69.8	80.9	64.6	79.7	88.3	77.2	48.6	50.7	33.3	–	–	–	61.2	59.3	45.7	–
	MiCE (ICLR20)	63.5	75.2	57.5	73.7	83.5	69.8	43.6	44.0	28.0	–	–	–	42.3	43.9	28.6	–
	CC (AAAI21)	76.4	85.0	72.6	70.5	79.0	63.7	43.1	42.9	26.6	85.9	89.3	82.2	44.5	42.9	27.4	62.1
	GCC (ICCV21)	68.4	78.8	63.1	76.4	85.6	72.8	47.2	47.2	30.5	84.2	90.1	82.2	49.0	52.6	36.2	64.3
	NNM (CVPR21)	66.3	76.8	59.6	73.7	83.7	69.4	48.0	45.9	30.2	–	–	–	60.4	58.6	44.9	–
	TCC (NeurIPS21)	73.2	81.4	68.9	79.0	90.6	73.3	47.9	49.1	31.2	84.8	89.7	82.5	55.4	59.5	41.7	67.2
	TCL (IJCV22)	79.9	86.8	75.7	81.9	88.7	78.0	52.9	53.1	35.7	87.5	89.5	83.7	62.3	64.4	51.6	71.4
	SPICE (TIP22)	81.7	90.8	81.2	73.4	83.8	70.5	44.8	46.8	29.4	82.8	92.1	83.6	57.2	64.6	47.9	68.7
	SIC (AAAI23)	95.3	98.1	95.9	**84.7**	**92.6**	**84.4**	59.3	58.3	43.9	97.0	98.2	96.1	69.0	69.7	55.8	79.9
	SeCu (ICCV23)	70.7	81.4	65.7	79.9	88.5	78.2	51.6	51.6	36.0	-	-	-	-	-	-	-
	DivClust (CVPR23)	-	-	-	71.0	81.5	67.5	44.0	43.7	28.3	85.0	90.0	81.9	51.6	52.9	37.6	-
	RPSC (AAAI24)	83.8	92.0	83.4	75.4	85.7	73.1	47.6	51.8	34.1	83.0	92.7	85.8	55.2	64.0	46.5	70.2
	LFSS (ICML25)	77.1	86.1	74.0	87.2	93.4	86.8	59.9	58.7	43.5	85.6	93.2	85.7	61.7	69.1	53.3	74.3
Training-free	CLIP (k-means)	91.7	94.3	89.1	70.3	74.2	61.6	49.9	45.5	28.3	96.9	98.2	96.1	39.8	38.1	20.1	66.3
	TAC (ICML24)	92.3	94.5	89.5	80.8	90.1	79.8	60.7	55.8	42.7	97.5	98.6	97.0	75.1	75.1	63.6	79.5
	CAE (Ours)	**95.5**	**98.2**	**96.7**	81.7	90.9	80.5	**62.8**	**60.9**	**45.9**	**98.1**	**98.8**	**97.4**	**77.9**	**77.5**	**66.3**	**81.9**
	CLIP (zero-shot)	93.9	97.1	93.7	80.7	90.0	79.3	55.3	58.3	39.8	95.8	97.6	94.9	73.5	72.8	58.2	78.7

Table 1: Clustering performance on 5 widely-used image clustering datasets. The best and second best results are denoted in **bold** and underline, respectively.

Method		DTD			UCF-101			ImageNet-1K			AVG
		NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
Training-based	SCAN (ECCV20)	59.4	46.4	31.7	79.7	61.1	53.1	74.7	44.7	32.4	53.7
	SIC (AAAI23)	59.6	45.9	30.5	81.0	61.9	53.6	77.2	47.0	34.3	54.6
Training-free	CLIP (k-means)	57.3	42.6	27.4	79.5	58.2	47.6	72.3	38.9	27.1	50.1
	TAC (ICML24)	60.1	45.9	29.0	81.6	61.3	52.4	77.8	48.9	36.4	54.8
	CAE (Ours)	**62.1**	**46.7**	**31.9**	**82.8**	**63.5**	**54.9**	**79.3**	**53.1**	**39.3**	**57.0**
	CLIP (zero-shot)	56.5	43.1	26.9	79.9	63.4	50.2	81.0	63.6	45.4	56.7

Table 2: Clustering performance on 3 challenging image clustering datasets. The best and second best results are denoted in **bold** and underline, respectively.

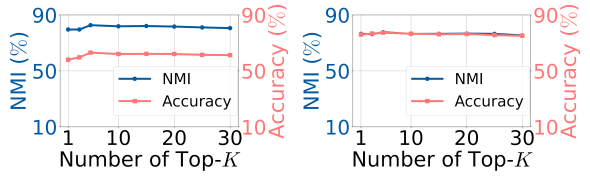
2020), IDFD (Tao, Takagi, and Nakata 2020), Mice (Tsai, Li, and Zhu 2020), CC (Li et al. 2021), GCC (Zhong et al. 2021), NNM (Dang et al. 2021), TCC (Wu et al. 2019), SeCu (Qian 2023), DivClust (Metaxas, Tzimiropoulos, and Patras 2023), PRSC (Liu et al. 2024), LFSS (Li et al. 2025). Our work mainly centers on comparison with zero-shot CLIP and CLIP-based methods. The results in Table 1 clearly show that our method consistently outperforms TAC (Li et al. 2024) (training-free) on 5 classic datasets, achieving a notable 7.2% improvement in ARI on STL-10 and a 5.1% accuracy improvement on CIFAR-20. This successful performance is attributed to the leveraged complementary semantics from selected nouns and captions, which demonstrate the effectiveness of our method. On the other hand, we observe that on the CIFAR-10 dataset, SIC (Cai et al. 2023) achieves the best performance, supported by the adaptation of features extracted from CLIP, which requires trainable parameters and additional training time. In contrast, our method is training-free and achieves the second-best results. Table 2 depicts the results on three challenge datasets, and our method still achieves the best performance. Specifically, our **CAE** outperforms TAC (Li et al. 2024) over 4% ACC on Imagenet-1K, which is a significant improvement. Moreover, the last row in Table 2 indicates the zero-shot performance of CLIP, which relies on the candidate class names of the images. Our method which depends solely on selected nouns and captions, outperforms zero-shot CLIP on DTD and UCF-101 datasets, highlighting the effectiveness of our approach in applying CLIP for clustering tasks.

	$\mathbf{x}^u$	$\mathbf{x}^v$	$\bar{\mathbf{x}}$	ImageNet-Dogs			UCF-101		
				NMI	ACC	ARI	NMI	ACC	ARI
(1)	✔			75.3	75.2	63.9	81.7	61.5	52.7
(2)		✔		75.7	75.5	64.0	81.4	61.3	52.3
(3)	✔	✔		76.4	76.3	64.5	82.3	62.1	52.9
(4)			✔	**77.9**	**77.5**	**66.3**	**82.8**	**63.5**	**54.9**

Table 3: Clustering performance of our method using different textual semantics.

Method	ImageNet-Dogs			UCF-101			AVG
	NMI	ACC	ARI	NMI	ACC	ARI	
(1) Concat($\mathbf{x}$, $\mathbf{x}^u$, $\mathbf{x}^v$)	75.7	75.6	64.2	79.7	59.4	50.2	53.6
(2) Sum($\mathbf{x}$, $\mathbf{x}^u$, $\mathbf{x}^v$)	76.3	76.2	65.6	82.3	62.1	52.9	55.6
(3) $\sum_{m\in\mathcal{M}} \beta_i^{(m)} \mathbf{x}_i^{(m)}$	**77.9**	**77.5**	**66.3**	**82.8**	**63.5**	**54.9**	**57.0**

Table 4: Clustering performance of our method with different combination strategies.



(a) UCF-101. (b) ImageNet-Dogs.

Figure 4: Analysis of clustering performance by varying the number of image semantic centers on (a) UCF-101 and (b) ImageNet-Dogs datasets, respectively.

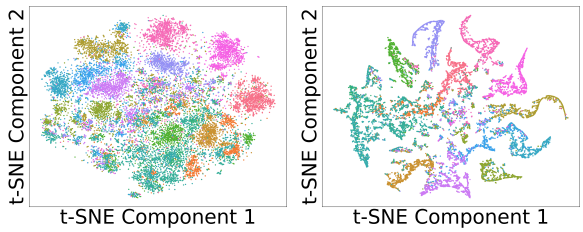
4.3 Ablation Study

We conduct experiments to assess the effectiveness of leveraged semantics and combination strategies. Additionally, we also examine the impact of the number of image semantic centers, and the number of top-K selections.

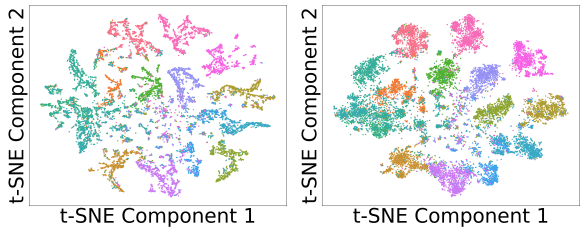
Effectiveness of leveraged semantics. Table 3 presents the results of using different textual semantics with images. In Row (1), when only using $\mathbf{x}^u$, it achieves comparable performance with that of Row (2), which uses only $\mathbf{x}^v$. When both $\bar{\mathbf{x}}^u$ and $\bar{\mathbf{x}}^v$ are combined in Row (3), there is a noticeable performance improvement, indicating that combining both semantic cues can better capture the image semantics. Finally, Row (4), which employs fused feature $\bar{\mathbf{x}}$, achieves the best results across all metrics and datasets. For instance, on the ImageNet-Dogs dataset, our method achieves a 2.6% improvement in NMI, a 2.3% improvement in ACC, and a 2.4% increase in ARI compared to Row (1).

Effectiveness of combination strategies. As discussed in Section 3.3, TAC directly concatenates noun embeddings with image embeddings. In contrast, our method adds both noun embeddings and captions embeddings to the image embeddings. The results in Table 4 demonstrate that the summation operation is more effective for image clustering compared to concatenation. Specifically, when comparing Rows (3), the adaptive fusion operation results in a 3.9% ACC improvement on the ImageNet-1K dataset using $\bar{\mathbf{x}}^u$ and $\bar{\mathbf{x}}^v$. These improvements suggest that our method better utilizes the semantics of both image and text, thereby enhancing the representativeness and discriminability of the embeddings.

Effects of top-K selection. In Eq. (3), we select the top-K nouns and descriptions of each image center to construct the semantic space. To examine the impact of the selection size, we vary top-K from 1 to 30 and evaluate the performance. As shown in Figure 4a and Figure 4b, using only one component (top-K = 1) fails to capture sufficient semantics, resulting in suboptimal clustering. On UCF-101 (Figure 4a), both NMI and ACC improve as top-K increases, stabilizing when a moderate number of components are used. However, further increasing top-K yields diminishing returns. A similar trend is observed on ImageNet-Dogs (Figure 4b), where excessive components introduce noise and slightly degrade performance. This highlights the need to balance semantic richness and redundancy for optimal clustering.



(a) CLIP Image Embeddings. (b) Nouns Embeddings ($\mathbf{x}^u$).



(c) Caption Embeddings ($\mathbf{x}^v$). (d) Our method (**CAE**).

Figure 5: Visualization of embeddings used for clustering on the ImageNet-Dogs dataset. a) image embeddings from CLIP. b) noun embeddings by Eq. (7). c) description embeddings by Eq. (8). d) embeddings by our method.

4.4 Visualization

We perform t-SNE (Van der Maaten and Hinton 2008) visualizations on features computed in our approach. Figure **??** shows that CLIP image embeddings exhibit only partial separation with notable overlaps. Figures 5b and 5c display noun ($\mathbf{x}^u$) and caption ($\mathbf{x}^v$) embeddings, both offering clearer separation, indicating that semantic cues improve image representation. Finally, Figure 5d illustrates our fused embeddings, where clusters become well-separated with minimal overlap, demonstrating markedly enhanced discriminability.

5 Conclusion

In this work, we propose a hierar**C**hical sem**A**ntic alignm**E**nt method for image clustering, dubbed **CAE**, which improves clustering performance without training. Our method constructs the semantic space by measuring distances between images and selected nouns and captions. Then, we adaptively fuse image, noun, and caption features according to their similarity with a semantic prototype. Extensive experiments demonstrate the effectiveness of our approach. In future work, we aim to obtain more precise descriptions for images by leveraging multimodal large language models.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No.U24B20180, No. 62576330), National Natural Science Foundation of Anhui (No.2508085MF143), and the advanced computing resources provided by the Supercomputing Center of the USTC.

References

Cai, S.; Qiu, L.; Chen, X.; Zhang, Q.; and Chen, L. 2023. Semantic-Enhanced Image Clustering. In *AAAI*.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep adaptive image clustering. In *ICCV*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607. PMLR.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *CVPR*, 3606–3613.

Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*.

Conti, A.; Fini, E.; Mancini, M.; Rota, P.; Wang, Y.; and Ricci, E. 2023. Vocabulary-free Image Classification. In *NeurIPS*.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *NeurIPS*.

Dang, Z.; Deng, C.; Yang, X.; Wei, K.; and Huang, H. 2021. Nearest neighbor matching for deep clustering. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Fang, A.; Ilharco, G.; Wortsman, M.; Wan, Y.; Shankar, V.; Dave, A.; and Schmidt, L. 2022. Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP). In *ICML*.

Gao, Q.; Lian, H.; Wang, Q.; and Sun, G. 2020. Cross-Modal Subspace Clustering via Deep Canonical Correlation Analysis. In *AAAI*.

Ge, Y.; Ren, J.; Gallagher, A.; Wang, Y.; Yang, M.; Adam, H.; Itti, L.; Lakshminarayanan, B.; and Zhao, J. 2023. Improving Zero-shot Generalization and Robustness of Multi-Modal Models. In *CVPR*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

Hu, W.; Miyato, T.; Tokui, S.; Matsumoto, E.; and Sugiyama, M. 2017. Learning discrete representations via information maximizing self-augmented training. In *ICML*.

Huang, J.; Gong, S.; and Zhu, X. 2020. Deep Semantic Clustering by Partition Confidence Maximisation. In *CVPR*.

Ji, P.; Zhang, T.; Li, H.; Salzmann, M.; and Reid, I. 2017. Deep Subspace Clustering Networks Pan. In *NeurIPS*.

Ji, X.; Henriques, J. F.; and Vedaldi, A. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*.

Li, Y.; Hu, P.; Liu, Z.; Peng, D.; Zhou, J. T.; and Peng, X. 2021. Contrastive clustering. In *AAAI*.

Li, Y.; Hu, P.; Peng, D.; Lv, J.; Fan, J.; and Peng, X. 2024. Image Clustering with External Guidance. In *ICML*.

Li, Y.; Yang, M.; Peng, D.; Li, T.; Huang, J.; and Peng, X. 2022. Twin Contrastive Learning for Online Clustering. *International Journal of Computer Vision*, 2205–2221.

Li, Z.; Jia, Y.; LIU, H.; and Hou, J. 2025. Learning from Sample Stability for Deep Clustering. In *ICML*.

Likas, A.; Vlassis, N.; and Verbeek, J. J. 2003. The global k-means clustering algorithm. *Pattern recognition*, 36(2): 451–461.

Liu, S.; Cao, W.; Fu, R.; Yang, K.; and Yu, Z. 2024. RPSC: Robust Pseudo-Labeling for Semantic Clustering. In *AAAI*.

Mao, Y.; Yan, X.; Guo, Q.; and Ye, Y. 2021. Deep Mutual Information Maximin for Cross-Modal Clustering. In *AAAI*.

Martin, S.; Huang, Y.; Shakeri, F.; Pesquet, J.; and Ayed, I. B. 2024. Transductive Zero-Shot and Few-Shot CLIP. In *CVPR*.

Menon, S.; and Vondrick, C. 2023. Visual Classification via Description from Large Language Models. In *ICLR*.

Metaxas, I. M.; Tzimiropoulos, G.; and Patras, I. 2023. DivClust: Controlling Diversity in Deep Clustering. In *CVPR*.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 39–41.

Niu, C.; Shan, H.; and Wang, G. 2022. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 7264–7278.

Novack, Z.; McAuley, J. J.; Lipton, Z. C.; and Garg, S. 2023. CHiLS: Zero-Shot Image Classification with Hierarchical Label Sets. In *ICML*.

Peng, X.; Feng, J.; Xiao, S.; Yau, W.-Y.; Zhou, J. T.; and Yang, S. 2018. Structured autoencoders for subspace clustering. *IEEE Transactions on Image Processing*, 5076–5086.

Peng, X.; Xiao, S.; Feng, J.; Yau, W.-Y.; and Yi, Z. 2016. Deep subspace clustering with sparsity prior. In *IJCAI*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *ICCV*.

Qian, Q. 2023. Stable Cluster Discrimination for Deep Clustering. In *ICCV*.

Qiu, J.; Wang, S.; Lu, J.; Liu, L.; Jiang, H.; Zhu, X.; and Hao, Y. 2025. Accelerating diffusion transformer via error-optimized cache. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 9588–9597.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Shaham, U.; and Stanton, K. 2018. SpectralNet: Spectral Clustering using Deep Neural Networks. In *ICLR*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR*.

Tang, Y.; Lin, Z.; Wang, Q.; Zhu, P.; and Hu, Q. 2024. AMU-Tuning: Effective Logit Bias for CLIP-based Few-shot Learning. In *CVPR*.

Tao, Y.; Takagi, K.; and Nakata, K. 2020. Clustering-friendly Representation Learning via Instance Discrimination and Feature Decorrelation. In *ICLR*.

Tsai, T. W.; Li, C.; and Zhu, J. 2020. Mice: Mixture of contrastive experts for unsupervised image clustering. In *ICLR*.

Van, G. W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. Scan: Learning to classify images without labels. In *ECCV*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9.

Villani, C.; et al. 2009. *Optimal transport: old and new*, volume 338. Springer.

Wu, J.; Long, K.; Wang, F.; Qian, C.; Li, C.; Lin, Z.; and Zha, H. 2019. Deep comprehensive correlation mining for image clustering. In *ICCV*.

Wu, Y.; Zhu, W.; Cao, J.; Lu, Y.; Li, B.; Chi, W.; Qiu, Z.; Su, L.; Zheng, H.; Wu, J.; and Yang, X. 2025. Video Repurposing from User Generated Content: A Large-scale Dataset and Benchmark. In *AAAI*, 8487–8495. AAAI Press.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*, 478–487.

Yang, B.; Fu, X.; Sidiropoulos, N. D.; and Hong, M. 2017. Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering. In *ICML*.

Yang, J.; Parikh, D.; and Batra, D. 2016. Joint unsupervised learning of deep representations and image clusters. In *CVPR*.

Yang, X.; Wu, Y.; Yang, M.; Chen, H.; and Geng, X. 2023. Exploring Diverse In-Context Configurations for Image Captioning. In *NeurIPS*.

Zhong, H.; Wu, J.; Chen, C.; Huang, J.; Deng, M.; Nie, L.; Lin, Z.; and Hua, X. 2021. Graph Contrastive Clustering. In *ICCV*.

Zhu, X.; Wang, S.; Lu, J.; Hao, Y.; Liu, H.; and He, X. 2024a. Boosting Few-Shot Learning via Attentive Feature Regularization. In *AAAI*.

Zhu, X.; Wang, S.; Zhu, B.; Li, M.; Li, Y.; Fang, J.; Wang, Z.; Wang, D.; and Zhang, H. 2025a. Dynamic Multimodal Prototype Learning in Vision-Language Models. *CoRR*, abs/2507.03657.

Zhu, X.; Zhang, R.; He, B.; Zhou, A.; Wang, D.; Zhao, B.; and Gao, P. 2023. Not All Features Matter: Enhancing Few-shot CLIP with Adaptive Prior Refinement. In *ICCV*.

Zhu, X.; Zhu, B.; Tan, Y.; Wang, S.; Hao, Y.; and Zhang, H. 2024b. Enhancing Zero-Shot Vision Models by Label-Free Prompt Distribution Learning and Bias Correcting. In *NeurIPS*.

Zhu, X.; Zhu, B.; Tan, Y.; Wang, S.; Hao, Y.; and Zhang, H. 2024c. Selective Vision-Language Subspace Projection for Few-shot CLIP. In *ACM Multimedia*, 3848–3857. ACM.

Zhu, X.; Zhu, B.; Wang, S.; Zhao, K.; and Zhang, H. 2025b. Enhancing CLIP Robustness via Cross-Modality Alignment. *arXiv preprint arXiv:2510.24038*.