Language-Assisted Image Clustering Guided by Discriminative Relational Signals and Adaptive Semantic Centers

Jun Ma[1], **Xu Zhang**[1], **Zhengxing Jiao**[1], **Yaxin Hou**[1], **Hui Liu**[3], **Junhui Hou**[4], **Yuheng Jia**[1,2,3*]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[3]School of Computing Information Sciences, Saint Francis University, Hong Kong, China
[4]Department of Computer Science, City University of Hong Kong, Hong Kong, China
220242351@seu.edu.cn, xuz@seu.edu.cn, jiaozx@seu.edu.cn, yaxin@seu.edu.cn
h2liu@sfu.edu.hk, jh.hou@cityu.edu.hk, yhjia@seu.edu.cn

Abstract

Language-Assisted Image Clustering (LAIC) augments the input images with additional textual information with the help of vision-language models (VLMs) to improve clustering performance. Despite recent progress, existing LAIC methods mainly construct image-text pairs for text-modality utilization, which overlook two issues: (i) textual features constructed for images are highly similar, leading to weak inter-class discriminability; (ii) the training is restricted to pre-built image-text alignments, limiting the potential for better utilization of the text modality. To address these issues, we propose a new LAIC framework with two complementary components. First, we exploit cross-modal relations to generate more discriminative self-supervision signals for clustering, which is compatible with the pre-training mechanisms of VLMs such as CLIP. Second, we learn category-wise continuous semantic centers via prompt learning to produce the final clustering assignments. Extensive experiments on eight benchmark datasets demonstrate that our method achieves an average improvement of **2.7%** over state-of-the-art methods, and the learned semantic centers exhibit strong interpretability. **Code is available at https://github.com/Mm077/LAIC-DRS-ASC.**

1 Introduction

Image clustering aims to partition images into semantically coherent clusters without supervision, with its core lying in mining reliable self-supervised signals. Existing deep image clustering methods [1, 2, 3, 4, 5] typically construct self-supervised signals within the image modality, and the quality of the obtained signals is inevitably constrained by the intrinsic information capacity of a single modality. Recently, by leveraging the strong image-text alignment capability of vision-language models (VLMs, e.g., CLIP [6]), **Language-Assisted Image Clustering (LAIC)** methods [7, 8, 9, 10] retrieve textual descriptions from external corpus (e.g., WordNet [11]) in unsupervised settings to view images in a multi-modal manner. These methods enhance image clustering performance, as textual semantics can provide complementary self-supervised signals for separating visually similar categories.

The key to LAIC design lies in effectively mining self-supervised information from the text modality to fully exploit textual semantics. Existing LAIC methods mostly adopt an **instance-wise pair-centric** paradigm for text-modality utilization, where the introduced texts are used to construct a textual

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

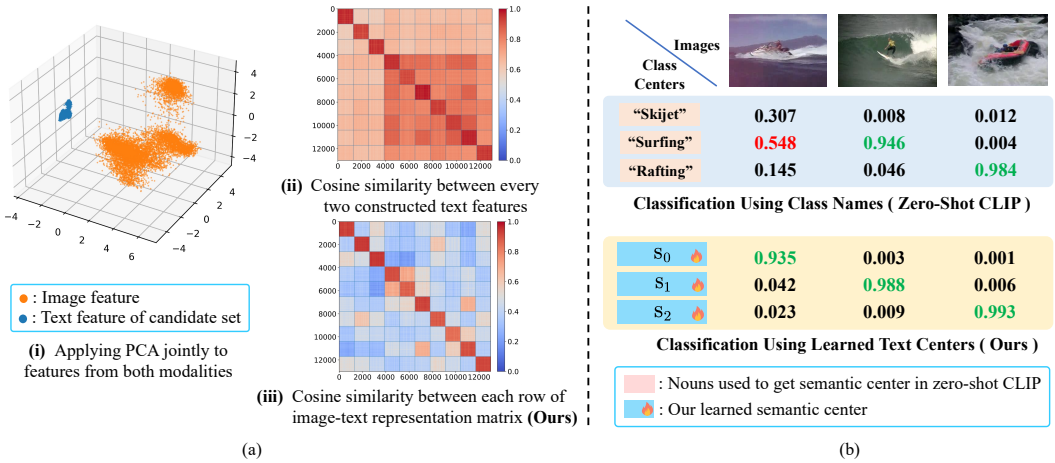


Figure 1: **(a)** Analysis of the phenomena of existing LAICs and our method on the ImageNet-10 dataset. (i) Joint PCA results show text features of candidate set exhibit a more compact distribution, which are closer to each other compared to image features. (ii) Due to the phenomenon in (i), the text features across different samples are overall highly similar, leading to weak inter-class discriminability. (iii) Rows of our learned image-text representation matrix serve as new representation, showing better inter-class discriminability with intra-class consistency retained. **(b)** Classification scores of three images (from UCF101 dataset) with respect to different semantic centers obtained by different methods. Zero-shot CLIP relies on semantic centers using ground-truth class names, which may not always capture accurate semantics (e.g, an image of motorboat is incorrectly assigned to "Surfing". Our learned semantic centers can even outperform those derived from ground-truth class names, showing stronger discriminability.

counterpart for each image. Specifically, they typically match several nearest words from candidate texts for each image, e.g., based on cosine similarity, construct image-text pairs, and exploit these pairs to mine self-supervised signals and train the clustering head for sample assignment.

Despite encouraging progress, we find that the existing **pair-centric** paradigm suffers from two main problems. **(i)** The text features matched to different samples are overall highly similar, which largely affects the discriminability of the supervision signals derived from the text modality (shown in Fig. 1(a-ii)). This likely occurs because the prompts used for the text encoder typically capture only coarse, high-level semantics, whereas images contain rich visual details and fine-grained information (e.g., an image of a cat playing on grass under a blue sky with clouds may be described simply as "a photo of a cat"). Thus, the text features of the candidate set are much closer to each other than image features (shown in Fig. 1(a-i)), leading to similar text features constructed for each sample. **(ii)** The current methods train the clustering model on existing manually constructed semantic space, which is constrained by previously established image-text alignments. This approach restricts the possibility of further extracting more accurate semantics from the native semantic space of VLMs.

To address these issues, we propose a new **global-wise relation-centric** LAIC framework. First, to enhance the discriminability of the text modality, we regard all candidate texts as semantic anchors in the textual space. We model global relations between images and these anchors rather than explicitly matching nearest nouns in a local manner. Specifically, we construct an image-text relational matrix by minimizing the discrepancy between the two modalities. Each row of this matrix characterizes how an image can be described from the perspective of the textual space, which can be treated as a new representation of this image. This approach is **compatible with the training mechanisms of VLMs** such as CLIP, where discriminability across different categories is largely driven by cross-modal information rather than solely text counterparts. Therefore, these new representations substantially improve inter-class discriminability and maintain intra-class compactness (shown in Fig. 1(a-iii)). We directly perform K-means clustering on the representation matrix to obtain a strong baseline.

Second, to move beyond the constraints of the constructed image-text matching relationships, we construct **learnable category-wise semantic centers** via prompt learning from the native semantic space of VLMs. Specifically, we initialize semantic centers by prompting the text encoder with the

template "a photo of a [class]", where [class] is replaced by a learnable vector for each pseudo-label class. We optimize these centers by maximizing the alignment between each pseudo-label class's semantic center and image features. The final clustering is achieved by assigning each sample to the closest semantic center. It is worth noting that, unlike the approaches in model adaptation where the prompt prefix is learned with class names available, we fix the prefix and optimize the semantic centers. Our approach closely mirrors CLIP's zero-shot process, and the results show that we can even learn better semantic centers than those constructed using the true class names (shown in Fig. 1(b)), suggesting that these centers can provide more accurate semantics from the perspective of the textual space. Meanwhile, the learned semantic centers exhibit strong interpretability.

Our contributions are summarized as follows:

- We propose a novel cross-modal relational mining method to utilize text modality which is highly compatible with the mechanisms of VLMs. By leveraging the inherent connections between image and text modalities, the constructed representations for samples are better in inter-class discriminability, providing richer self-supervision signals and a strong baseline for clustering.
- We introduce a continuous category-wise semantic centers learning strategy by prompt learning directly in the semantic space of VLMs. This approach moves beyond the constraints of the pre-built image-text alignments and acquires more accurate semantic centers, allowing more effective utilization of the text modality.
- Experimental results on eight public datasets demonstrate that our method achieves an average improvement of **2.7%** over state-of-the-art methods, and the learned semantic centers exhibit strong interpretability.

2 Related Work

2.1 Deep Image Clustering

Traditional image clustering mainly relies on K-means [12] and Spectral Clustering [13], which struggle to handle high-dimensional data. Methods such as DEC [14] and DeepCluster [15] introduce representation learning into the clustering pipeline, marking the rise of deep clustering. Approaches including SCAN [1] and SPICE [16] leverage pseudo-labeling to further improve clustering by selecting high-confidence pseudo-labels from the obtained cluster assignments. Meanwhile, contrastive learning has contributed to notable progress in image clustering. CC [17] and TCL [18] perform contrastive learning at both instance and cluster levels, while ProPos [2] combines non-contrastive learning at the instance level with contrastive learning at the cluster level. Some methods analyze phenomena observed in clustering from different perspectives. CDC [19] identifies the overconfidence problem in deep clustering and alleviates it by introducing a model structure with calibration head. DCBoost [20] studies the inconsistency between global and local feature structures in deep clustering, and proposes a plug-in module that enhances global features by leveraging local structural information. With the advent of vision-language models (VLMs, e.g., CLIP [6]), methods such as TEMI [21] and CPP [22] exploit the strong representations learned by VLMs to enhance clustering performance.

2.2 Language-Assisted Image Clustering

Language-Assisted Image Clustering incorporates supervision signals extracted from the text modality to enhance image clustering performance, as textual semantics can provide additional guidance, especially when categories are visually similar but semantically different. SIC [7] derives pseudo-label supervision according to relationships between images and semantics to guide image clustering, while MCA [8] proposes a hierarchy-based nouns filtering strategy and improves image-text alignments at three levels. TAC [9] treats signals from text modality as external guidance, enhancing clustering performance via self-distillation between two modalities. GradNorm [10] selects more accurate texts that are more semantically aligned with images from the gradient perspective. Despite their differences, these methods explicitly match each image with several semantically nearest texts and construct textual counterparts for images. Such local text matching may provide limited discriminative supervision, since the matched textual features for different images can be highly similar. Moreover, the clustering process is largely constrained by the pre-built image-text alignments, leaving limited room to further explore more suitable semantics.

3 Method

Overview. Given a set of unlabeled images $\mathcal{X} = \{x_i\}_{i=1}^N$, image clustering aims to partition these samples into semantically coherent clusters, where only the number of clusters is known. To this end, we propose a new LAIC framework built upon CLIP with two components. First, we design a cross-modal relation mining scheme to extract discriminative self-supervision signals in Sec. 3.1. Second, we learn continuous semantic centers for each class from CLIP's semantic space in Sec. 3.2. These centers serve as discriminative anchors and produce the final clustering results.

3.1 Semantic Space Construction and Supervision Mining

Constructing dataset-specific semantic space. We adopt CLIP as the backbone and denote its image encoder and text encoder as $f(\cdot)$ and $g(\cdot)$, respectively. We extract image feature $\mathbf{x}_i$ for each $x_i \in \mathcal{X}$ using $f(\cdot)$ and form the image feature matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times d}$. Meanwhile, we encode each word w_i from the external corpus WordNet [11] into a text feature $\mathbf{w}_i$ using $g(\cdot)$ and form the text feature matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L]^\top \in \mathbb{R}^{L \times d}$, where L is the size of WordNet.

To obtain a condidate nouns set that can semantically describe the current dataset, we first compute fine-grained semantic centers of images. Specifically, we run K-means clustering on $\mathbf{X}$ to produce $\tilde{k}$ clusters. We set $\tilde{k} = \lceil N/300 \rceil$; for datasets where the number of samples in each ground-truth class is smaller than 300, we use $\tilde{k} = 3K$, where K is the number of clusters. Let $\{\mathcal{P}_r\}_{r=1}^{\tilde{k}}$ be the resulting partition. The r-th fine-grained semantic center is computed as the mean feature:

$$\mathbf{p}_r = \frac{1}{|\mathcal{P}_r|} \sum_{i \in \mathcal{P}_r} \mathbf{x}_i, \quad r = 1, \ldots, \tilde{k}. \tag{1}$$

Next, each text feature $\mathbf{w}_i$ from $\mathbf{W}$ is assigned to its nearest image center $\mathbf{p}_{a(i)}$, where

$$a(i) = \underset{r \in \{1, \ldots, \tilde{k}\}}{\operatorname{argmax}} \cos(\mathbf{w}_i, \mathbf{p}_r), \quad i = 1, \ldots, L, \tag{2}$$

where $\cos(a, b) = \frac{a^\top b}{\|a\|_2 \|b\|_2}$ denotes cosine similarity. Then, for each center $\mathbf{p}_r$, we select θ most similar text features among those assigned to it, noted by $\mathbf{U}_r$. Finally, we take the union $\bigcup_{r=1}^{\tilde{k}} \mathbf{U}_r$ as the dataset-specific text features candidate set $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_M]^\top \in \mathbb{R}^{M \times d}$.

Mining discriminative supervision from cross-modal relational signals. Existing LAIC methods typically retrieve text feature for each image based on the several nearest text features in $\mathbf{U}$ and construct supervision. However, as illustrated in Fig.1(a-ii), this strategy overlooks that the text features constructed for samples are highly similar overall. Moreover, it ignores abundant discriminative information carried by *non-neighbor* texts, as samples belonging to the same semantic category typically exhibit consistent dissimilarity with the same irrelevant texts.

To address these issues, we regard all candidate texts as semantic anchors in the textual space and model the global relations between each image and these anchors rather than explicitly matching nearest nouns in a local manner. Concretely, we mine supervision from a cross-modal perspective by learning an image-text representation matrix $\mathbf{C} \in \mathbb{R}^{N \times M}$. It reconstructs image features using the entire candidate set $\mathbf{U}$:

$$\min_{\mathbf{C}} \ \|\mathbf{X} - \mathbf{C}\mathbf{U}\|_F^2 + \gamma \|\mathbf{C}\|_F^2, \tag{3}$$

where the first term enforces that each image feature can be well approximated by a weighted combination of text features, and the Frobenius regularizer imposes a smoothness constraint on $\mathbf{C}$. γ is the regularization parameter. Problem (3) is a standard ridge regression with a closed-form solution:

$$\mathbf{C}^\star = \mathbf{X}\mathbf{U}^\top \left(\mathbf{U}\mathbf{U}^\top + \gamma \mathbf{I}_M\right)^{-1}, \tag{4}$$

where $\mathbf{I}_M \in \mathbb{R}^{M \times M}$ represents the identity matrix. Each row $\mathbf{c}_i$ of $\mathbf{C}^\star$ can be interpreted as a dense semantic description of image x_i over the whole candidate noun set. It becomes a discriminative representation for x_i, for semantically similar images tend to be compact in this space, and vice versa. In practice, we directly perform K-means on the rows of $\mathbf{C}^\star$ to obtain an initial clustering baseline, which also serves as improved supervision for the subsequent semantic center learning step.

Remark 1. *Why are cross-modal relational signals more discriminative?* Recall that CLIP aligns a large number of images with texts that describe their semantics during pretraining. Thus, CLIP's discriminative ability across different categories is largely driven by cross-modal information, i.e., the difference between feature similarities of images and texts. The cross-modal relational signals provided by $\mathbf{C}$ align well with this mechanism of CLIP.

3.2 Semantic Centers Learning

In this part, we learn semantic centers as continuous and discriminative cluster anchors in CLIP's semantic space, enabling stable test-time clustering without relying on a discrete candidate noun set.

High-quality pseudo-labels selection. We retain only the pseudo-label information from Sec. 3.1 to overcome the limitations of pre-built image-text alignments. Although the previous step provides a strong baseline, the pseudo-labels still contain noise. We therefore filter them to obtain high-quality pseudo-labels for more reliable supervision. To this end, we first define sample neighborhood relations in the space of $\mathbf{C} \in \mathbb{R}^{N\times M}$, where $\mathbf{c}_i$ denotes the i-th row of $\mathbf{C}$. We compute $\cos(\mathbf{c}_i, \mathbf{c}_j)$ as similarity between x_i and x_j, where $j \neq i$. For each sample x_i, we select the top-$\hat{k}$ most similar samples in space of $\mathbf{C}$ as its $\hat{k}$-nearest-neighbors ($\hat{k}$-nn) set $\mathcal{N}_{\hat{k}}(i)$.

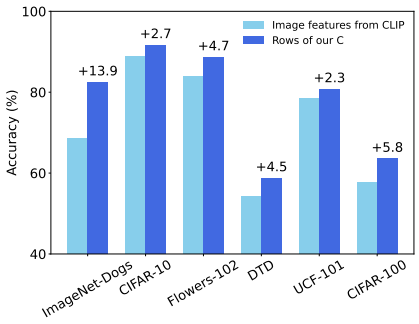


Figure 2: Accuracy of each sample and its $\hat{k}$-nn ($\hat{k} = 10$) belonging to the same ground-truth class in different spaces. Our representation achieves better neighbor consistency.

We observe that a sample and its $\hat{k}$-nn have a high probability of belonging to the same semantic category (shown in Fig. 2). Based on this observation, we hypothesize that *the stronger the pseudo-label consistency between a sample and its neighbors, the more likely the pseudo-label is correct*. Therefore, for $\hat{y}_i \in \{1, \ldots, K\}$, we define a neighbor-consistency score:

$$\alpha_i = \frac{1}{\hat{k}} \sum_{j\in\mathcal{N}_{\hat{k}}(i)} \mathbb{I}\big[\hat{y}_j = \hat{y}_i\big], \tag{5}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. Higher score means that $\hat{y}_i$ is more likely to be accurate. We select samples with $\alpha_i = 1$ to form the high-quality set $\mathcal{D}_L = \{\,(x_i, \hat{y}_i) \mid \alpha_i = 1\}$ as supervision for semantic centers learning.

Semantic centers parameterization. With the pre-trained parameters of CLIP frozen, we adopt prompt learning to adapt the text-side representations to the target data distribution. We replace the *class* position in the prompt "a photo of a [class]" with randomly initialized learnable vectors, which serve as the semantic-center variables of pseudo-label classes. For each class k, we introduce learnable class vectors $\mathbf{v}_k$, which is a continuous vector in the word embedding space. By concatenating these learnable vectors to fixed prefix vectors, we construct the prompt for class k as $\mathbf{t}_k = [\mathbf{E}_{prefix}, \mathbf{v}_k]$, where $\mathbf{E}_{prefix}$ contains the embeddings of the fixed prefix tokens, and we set prefix as "a photo of a". We feed $\mathbf{t}_k$ into the frozen text encoder $g(\cdot)$ of CLIP to obtain the semantic center's text feature $\mathbf{s}_k$ for class k, and $\mathcal{S} = \{\mathbf{s}_k\}_{k=1}^{K}$ denotes the set of learnable semantic centers.

Remark 2. *What is the difference between prompt tuning in model adaptation and our learning of semantic centers?* In field of model adaptation, the class names are available, and the approaches often replace the prompt prefix before *class* with learnable vectors. The goal is to learn prompts that are better adapted to downstream tasks, in order to better leverage the generalization ability of VLMs. With the unsupervised setting of LAIC, our method fixs the prefix and replace *class* with learnable vectors, aiming to learn continuous semantic centers to get clustering results.

Semantic centers training objective. Given an image x, we extract its image feature $\mathbf{x} = f(x)$, where $f(\cdot)$ is the frozen CLIP image encoder. We define the probability of assigning x to k as

$$p(k \mid x) = \frac{\exp(\ell_k(x))}{\sum_{j=1}^{K} \exp(\ell_j(x))}, \tag{6}$$

where $\ell_k(x) = \cos(\mathbf{x}, \mathbf{s}_k)/T$ and T is the temperature parameter learned during pre-training of CLIP. Thus, the predicted probability distribution of x over K classes is

$$\mathbf{P}(x) = \big(p(1 \mid x), \ldots, p(K \mid x)\big). \tag{7}$$

We set $\mathcal{D}_L = \{(x_i, \hat{y}_i) \mid \alpha_i = 1\}$ as the high-quality set, where $\hat{y}_i \in \{1, \ldots, K\}$ is pseudo-label, and $\mathcal{D}_U = \{x_i \mid \alpha_i < 1\}$ as the unlabeled set.

First, on $\mathcal{D}_L$, we minimize the generalized cross-entropy (GCE) as supervised loss to enhance the model's robustness against noise in pseudo-labels. The loss is formulated as

$$\mathcal{L}_{\text{sup}} = \mathbb{E}_{(x,\hat{y})\sim\mathcal{D}_L} \left[\frac{1 - p\big(\hat{y} \mid \mathcal{A}_s(x)\big)^q}{q} \right], \tag{8}$$

where $\mathcal{A}_s(\cdot)$ denotes the strong augmentation, $q \in (0, 1]$ balances the convergence speed of cross-entropy loss and the noise robustness of mean absolute error loss. We set $q = 0.8$. This term encourages samples from $\mathcal{D}_L$ to be close to the semantic center of their assigned class guided by pseudo-labels and far from other centers, thus improving inter-class separability.

Next, to provide stable constraints on unlabeled data, we generate weak augmentation for each $x \in \mathcal{D}_U$, denoted as $\mathcal{A}_w(x)$. We enforce prediction consistency between the two views by minimizing the mean squared error between the two probability distributions:

$$\mathcal{L}_{\text{con}} = \mathbb{E}_{x\sim\mathcal{D}_U} \left[\big\| \mathbf{P}\big(\mathcal{A}_s(x)\big) - \mathbf{P}\big(\mathcal{A}_w(x)\big) \big\|_2^2 \right]. \tag{9}$$

Then, to encourage a sufficiently diverse usage of the K centers, we maximize a global entropy regularizer:

$$\mathcal{L}_{\text{ent}} = -\sum_{k=1}^{K} q(k) \log q(k), \tag{10}$$

where we define $q(k) = \mathbb{E}_{x\sim(\mathcal{D}_L \cup \mathcal{D}_U)} \big[p(k \mid x)\big]$ as the global average prediction distribution on both high-quality set and unlabeled set. Finally, we combine the above terms and get the overall objective:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_1 \mathcal{L}_{\text{con}} - \lambda_2 \mathcal{L}_{\text{ent}}, \tag{11}$$

where weight parameters $\lambda_1 = 2$ and $\lambda_2 = 0.1$ are fixed on all datasets.

3.3 Model Training and Prediction

First, we extract pseudo-labels as discriminative self-supervision signals from text modality by Eq. (4). Then we freeze CLIP's image and text encoders and only update $\mathcal{V} = \{\mathbf{v}_k\}_{k=1}^K$ in semantic centers using Eq. (11). After training, we obtain K semantic centers $\mathcal{S} = \{\mathbf{s}_k\}_{k=1}^K$. For a test image x, we extract its image feature $\mathbf{z} = f(x)$ and assign it to the closest semantic center:

$$\tilde{y}(x) = \underset{k}{\text{argmax}} \ \cos(\mathbf{z}, \mathbf{s}_k). \tag{12}$$

We partition images with the same $\tilde{y}$ to the same cluster and get final cluster results.

4 Experiments

4.1 Experimental Setup

We apply our method to four widely-used image clustering datasets including ImageNet-Dogs [33], ImageNet-10 [33], STL-10 [34], CIFAR-10 [35], along with four datasets with larger cluster numbers, including Flowers-102 [36], CIFAR-100 [35], DTD [37] and UCF-101 [38]. We compare our method with 18 deep clustering approaches and zero-shot CLIP. We evaluate clustering performance using NMI, ACC, and ARI, where higher values indicate better performance. We adopt the pre-trained CLIP model with ViT-B/32 [39] and Transformer [40] as image and text backbones, respectively. We fix $\theta = 2$, $\gamma = 5$, and $\hat{k} = 10$ for all experiments, except for datasets with larger cluster numbers, where $\hat{k} = 1$. All hyperparameters are selected without using the test set or any tuning procedure, and our method remains robust to their values within a reasonable range (seen in Appendix E). More implementation details are provided in the Appendix A and B.

Table 1: Clustering performance (%) on four widely used datasets. ZS-CLIP means zero-shot CLIP.

Datasets		ImageNet-Dogs			ImageNet-10			STL-10			CIFAR-10			Avg.
Methods	Backbone	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
DEC [14]	ResNet	12.2	19.5	7.9	28.2	38.1	20.3	13.6	18.5	5.0	25.7	30.1	16.1	19.6
IIC [23]	ResNet	-	-	-	-	-	-	49.6	59.6	39.7	51.3	61.7	41.1	-
DCCM [24]	ResNet	32.1	38.3	18.2	60.8	71.0	55.5	37.6	48.2	26.2	49.6	62.3	40.8	45.1
PICA [25]	ResNet	33.6	32.4	17.9	78.2	85.0	73.3	59.2	69.3	50.4	56.1	64.5	46.7	55.6
BYOL [26]	ResNet	69.7	72.9	60.9	88.4	94.7	88.9	75.4	86.1	71.5	78.0	87.5	75.2	79.1
SCAN [1]	ResNet	61.2	59.3	45.7	-	-	-	69.8	80.9	64.6	79.7	88.3	77.2	-
CC [17]	ResNet	44.5	42.9	27.4	85.9	89.3	82.2	76.4	85.0	72.6	70.5	79.0	63.7	68.3
MiCE [27]	ResNet	42.3	43.9	28.6	-	-	-	63.5	75.2	57.5	73.7	83.5	69.8	-
GCC [28]	ResNet	49.0	52.6	36.2	84.2	90.1	82.2	68.4	78.8	63.1	76.4	85.6	72.8	70.0
TCC [29]	ResNet	55.4	59.5	41.7	84.8	89.7	82.5	73.2	81.4	68.9	79.0	90.6	73.3	73.3
TCL [18]	ResNet	62.3	64.4	51.6	87.5	89.5	83.7	79.9	86.8	75.7	81.9	88.7	78.0	77.5
DMICC [30]	ResNet	58.1	58.7	43.8	91.7	96.2	91.6	68.9	80.0	62.5	74.0	82.8	69.0	73.1
LFSS [31]	ResNet	61.7	69.1	53.3	85.6	93.2	85.7	77.1	86.1	74.0	84.1	92.4	84.2	78.9
SIC [7]	ViT-B/32	69.0	69.7	55.8	97.0	98.2	96.1	95.3	98.1	95.9	84.7	92.6	84.4	86.4
MCA [8]	ViT-B/32	73.3	74.9	61.6	-	-	-	95.5	98.1	96.0	84.9	92.7	84.6	-
TAC [9]	ViT-B/32	80.6	83.0	72.2	98.5	99.2	98.3	95.5	98.2	96.1	83.3	91.9	83.1	90.0
PRO-DSC [32]	ViT-B/32	-	-	-	98.0	99.0	97.8	95.4	98.1	95.9	79.6	87.1	80.2	-
GradNorm [10]	ViT-B/32	81.0	81.2	70.9	98.7	99.4	98.7	95.6	98.3	96.2	82.6	91.1	81.5	89.6
Ours	ViT-B/32	**86.2**	**88.0**	**82.6**	**99.6**	**99.8**	**99.7**	**96.1**	**98.5**	**96.7**	**85.2**	**92.9**	**84.8**	**92.5**
ZS-CLIP [6]	ViT-B/32	73.5	72.8	58.2	95.8	97.6	94.9	93.9	97.1	93.7	80.7	90.0	79.3	85.6

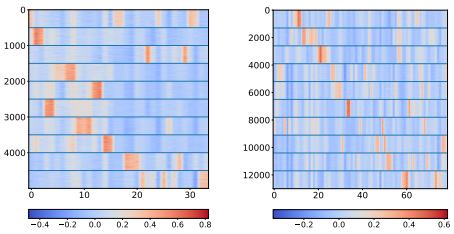
Table 2: Clustering performance (%) on four datasets with larger cluster numbers.

Datasets		Flowers-102			CIFAR-100			DTD			UCF-101			Avg.
Methods	Backbone	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
SCAN [1]	ResNet	77.9	56.5	50.9	55.4	38.7	25.1	59.4	46.4	31.7	79.7	61.1	53.1	53.0
LFSS [31]	ResNet	79.2	55.4	51.8	66.0	52.4	37.6	58.7	45.5	31.9	58.9	34.5	21.6	49.5
SIC [7]	ViT-B/32	86.1	69.2	62.7	63.5	48.3	34.7	59.6	45.9	30.5	81.0	61.9	52.4	58.0
TAC [9]	ViT-B/32	87.1	70.2	63.1	68.1	55.4	42.4	62.1	50.1	34.4	82.8	68.9	60.6	62.1
PRO-DSC [32]	ViT-B/32	83.2	69.8	59.6	68.2	55.7	40.2	57.6	48.1	31.3	81.0	61.9	52.8	59.1
GradNorm [10]	ViT-B/32	-	-	-	67.2	54.8	37.0	63.1	50.9	34.2	82.9	63.2	53.9	-
Ours	ViT-B/32	**88.6**	**76.6**	**73.9**	**68.8**	**58.1**	**43.4**	**63.8**	**53.7**	**36.2**	**84.4**	**70.2**	**61.9**	**65.0**
ZS-CLIP [6]	ViT-B/32	82.3	66.7	58.1	67.4	60.9	38.0	56.5	43.1	26.9	79.9	63.4	50.2	57.8

4.2 Main Results

Clustering performance. We report results on four widely used datasets in Table 1 and four datasets with larger cluster numbers in Table 2. Best and second-best results are highlighted in bold and underline, respectively. We focus on comparisons with CLIP-based methods and zero-shot CLIP. Our method consistently outperforms all baselines, achieving an average improvement of **2.7%** over the state-of-the-art LAIC method. The gains are more pronounced on datasets with larger cluster numbers, where our method improves ACC by an average of **3.5%** over TAC. On fine-grained datasets, the improvements are particularly significant: **+5.0%** ACC and **+10.4%** ARI on ImageNet-Dogs, and **+6.4%** ACC and **+10.8%** ARI on Flowers-102. We also outperform zero-shot CLIP by an average of **7.0%**, demonstrating the effectiveness of our learned semantic centers.

Visualization of image-text representation matrix. To better demonstrate that the new representations $\mathbf{C}$ learned in Sec. 3.1 are more discriminative, we visualized the entire $\mathbf{C}$ on two datasets in Fig. 3. Rows associated with images of the same class are grouped between two adjacent horizontal lines. Within each class, the high and low value regions of different rows show clear consistency, indicating strong intra-class similarity; between different classes, the high and low value regions exhibit distinct differences, reflecting inter-class discriminability.



(a) STL-10 (b) ImageNet-10

Figure 3: Visualization of $\mathbf{C}$

Interpretability of learned semantic centers. To better illustrate the interpretability of the learned semantic centers in Sec. 3.2, we selected the nearest nouns for each learned centers from the candidate noun set and compare them with ground-truth class names in Table 3. As can be shown, the selected nouns for learned semantic centers align well with the ground-truth semantic of the dataset, demonstrating strong interpretability. Therefore, these results show that our method not only achieves effective clustering but also provides accurate textual descriptions for each class, thereby improving the richness of the clustering outcomes.

Table 3: Comparisons between the most similar nouns for learned semantic centers and the corresponding ground-truth (GT) class names.

ImageNet-10		STL-10	
Most similar noun	GT class	Most similar noun	GT class
sports_car	**Sports Car**	draft_horse	**Horse**
airline	**Airliner**	tabby_cat	**Cat**
soccer_ball	**Soccer Ball**	container_ship	**Ship**
snow_leopard	**Snow Leopard**	sports_car	**Car**
king_penguin	**King Penguin**	riflebird	**Bird**
navel_orange	**Orange**	guenon_monkey	**Monkey**
Maltese_dog	**Maltese Dog**	whitetail_deer	**Deer**
Blimp	**Airship**	trucking_rig	**Truck**
trucking_rig	**Trailer Truck**	multiengine_airplane	**Airplane**
containership	**Container Ship**	domestic_dog	**Dog**

4.3 Analysis

Comprasions on different backbones. We validate the effectiveness of our method on more backbones, including CLIP (ViT-B/16) and BLIP [41](ViT-B/16). Experiments on three datasets show that our supervision signal extraction strategy and final clustering results remains better than existing methods, demonstrating the generalizability of our observations and method.

Table 4: Clustering performance (in percent %) across three datasets with more backbones.

Datasets		CIFAR-10			DTD-47			UCF-101			Avg.
Methods	Backbone	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
SIC [7]	CLIP ViT-B/16	85.7	93.2	85.9	61.9	51.0	35.8	77.6	65.7	54.5	67.9
TAC [9]	CLIP ViT-B/16	85.3	92.9	85.2	63.3	52.1	37.9	83.5	70.9	61.9	70.3
Ours	CLIP ViT-B/16	**87.9**	**94.5**	**88.2**	**65.1**	**55.2**	**39.7**	**85.2**	**71.8**	**63.1**	**72.3**
K-means	CLIP ViT-B/16	75.4	79.3	76.9	57.3	45.7	30.9	79.7	63.5	53.9	62.5
SIC [7]	BLIP ViT-B/16	82.3	91.5	81.7	62.5	53.0	38.4	75.4	59.8	50.3	66.1
TAC [9]	BLIP ViT-B/16	81.4	90.6	80.9	64.9	55.8	40.9	78.0	62.1	53.2	67.5
Ours	BLIP ViT-B/16	**84.2**	**93.5**	**84.4**	**66.8**	**57.0**	**42.8**	**80.2**	**64.9**	**56.8**	**70.1**
K-means	BLIP ViT-B/16	79.4	88.3	77.9	62.7	55.3	38.4	78.0	61.2	51.5	65.9

Effectiveness analysis of selecting high-quality pseudo-labels. We compare the accuracy between pseudo-labels by conducting K-means clustering on $\mathbf{C}$ and high-quality pseudo-labels after selecting. Results shown in Fig. 4 across eight diverse datasets demonstrate a substantial enhancement in pseudo-label accuracy following the neighborhood consistency filtering. Notably, significant gains are observed in datasets with lower performance on $\mathbf{C}$, such as ImageNet-Dogs (**+15.2%**), CIFAR-100 (**+10.1%**), and DTD (**+11.7%**). These improvements validate that local structural consistency serves as a robust instruction for identifying high-quality samples while effectively decreasing noise near cluster boundaries. Thus, selection of high-quality subset provides a cleaner supervisory instruction for the learning of semantic centers.

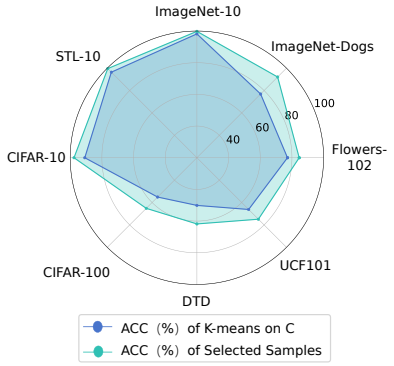


Figure 4: Accuracy gains after filtering.

Cross-domain generalization of learned semantic centers. To evaluate the generalization ability of the learned semantic centers, we conduct experiments on the ImageNet-R [42] and ImageNet-V2 [43] dataset (shown in Table 5). ImageNet-R contains 200 classes derived from a subset of ImageNet's 1,000 classes, while features a rendition of the ImageNet classes in di-

Table 5: Cross-domain generalization of learned semantic centers.

Datasets	IN-20			IN-R-20			IN-V2-20			Avg.
Methods	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
SIC [7]	92.8	93.7	90.1	70.6	82.5	71.3	66.2	78.4	70.9	79.6
TAC [9]	95.8	96.2	91.3	74.6	84.5	73.6	70.5	82.1	73.6	82.5
Ours	**96.9**	**97.2**	**92.8**	**78.9**	**88.9**	**79.5**	**75.7**	**84.9**	**76.9**	**85.7**

verse image styles such as paintings, cartoons and sculptures. ImageNet-V2 is a reproduced test set using different sources while following ImageNet's data collection process. Specifically, we select 20 classes from ImageNet (including strawberry, banana, cellphone and so on) to learn the semantic centers, and then test them as semantic anchors to get clustering results on both original test set (noted as IN-20 in Table 5) and the corresponding classes in cross-domain datasets (noted as IN-R-20 and IN-V2-20). Our learned semantic centers remain more discriminative and robust on a semantically similar yet distribution-shifted dataset. This demonstrates that the proposed method can capture more transferable textual semantics rather than overfitting to dataset-specific visual patterns, thereby achieving stronger generalization ability.

Robustness to dictionary construction strategies. To investigate the robustness of our method to dictionary construction, we evaluate two alternative noun selection strategies on STL-10 and DTD-47, with K-means performed on the resulting representation C. First, we randomly sample different numbers of nouns from WordNet to replace our semantic selection strategy. As shown in Table 6, increasing the dictionary size gradually improves performance, but even 2,000 randomly selected nouns remain inferior to our method, which uses only 31 and 278 nouns on STL-10 and DTD-47, respectively. Second, we adopt a match-frequency-based strategy that skips the initial image K-means. Specifically, we retrieve the top-r WordNet nouns for each image and select the most frequent nouns across all retrieval results, matching the dictionary size of our method. As shown in Table 7, this strategy achieves comparable but slightly lower performance, with larger r introducing potentially less relevant nouns. Overall, these results demonstrate that our method does not rely on highly refined noun selection, while our strategy provides a compact and effective candidate semantic space.

Table 6: Clustering accuracy (%) under different random dictionary sizes.

Dictionary Construction	#Nouns	STL-10	DTD-47
Random	100	88.7	39.4
	500	91.9	41.9
	1,000	93.1	44.1
	2,000	95.6	47.8
Proposed	31 / 278	**96.4**	**49.9**

Table 7: Clustering accuracy (%) under match-frequency-based dictionary construction.

Dictionary Construction	Retrieval Size r	STL-10	DTD-47
Match-frequency	1	96.0	49.2
	5	95.6	48.6
	10	94.9	47.8
Proposed	–	**96.4**	**49.9**

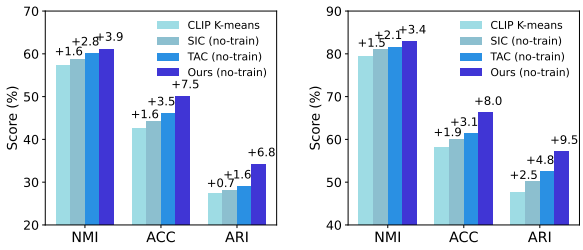
4.4 Ablation Study

Loss terms. To evaluate the effectiveness of the three loss terms $\mathcal{L}_{\text{sup}}$, $\mathcal{L}_{\text{con}}$, and $\mathcal{L}_{\text{ent}}$, we evaluate the performance of our method using different combinaions with these losses. From the results in Table 8, we observe that the supervised loss $\mathcal{L}_{\text{sup}}$ is crucial for establishing a solid clustering foundation, as evidenced by the performance when using only this term. Meanwhile, the consistency loss $\mathcal{L}_{\text{con}}$ and the entropy loss $\mathcal{L}_{\text{ent}}$ further enhance clustering performance, but their effects are limited without the supervised loss. Combining all three losses achieves the best performance, confirming that each loss term contributes to the overall effectiveness of our method.

Table 8: Different loss combinations.

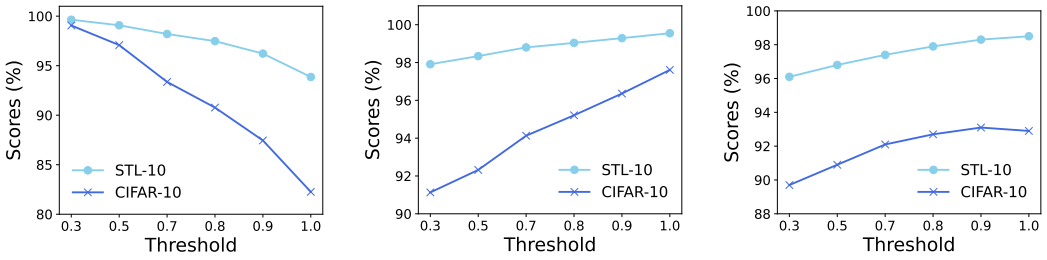
$\mathcal{L}_{\text{sup}}$	$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{ent}}$	ImageNet-Dogs			DTD		
			NMI	ACC	ARI	NMI	ACC	ARI
✓			82.4	81.7	75.2	60.5	51.8	34.1
	✓		8.1	14.3	3.6	10.2	12.4	8.6
		✓	9.1	14.6	5.7	13.5	16.2	9.7
✓	✓		86.0	87.6	81.7	62.6	51.9	34.3
✓		✓	85.9	87.4	82.1	63.3	53.2	35.5
	✓	✓	10.6	17.6	6.8	15.5	30.1	10.5
✓	✓	✓	**86.2**	**88.0**	**82.6**	**63.8**	**53.7**	**36.2**

Comparisons of different K-means baselines. To demonstrate that we have extracted more discriminative relational signals in Sec. 3.1, we directly perform K-means clustering on the image-text representation matrix $\mathbf{C}$ as a clustering baseline, referred to as Ours (no-train). We compare the performance of Ours (no-train) with three baseline methods, shown in Fig. 5. First, the CLIP baseline, which conducts K-means on image features from CLIP. Second, SIC (no-train) and TAC (no-train), which reports the pseudo-label performance obtained from the corresponding method. Compared to SIC (no-train) and TAC (no-train), our method shows a greater improvement over the CLIP baseline. The improvement over both baselines is most notable on the DTD dataset (+4.0% in ACC, +5.2% in ARI compared to TAC (no-train) and +7.5% in ACC, +6.8% in ARI compared to CLIP baseline).



(a) DTD (b) UCF-101

Figure 5: K-means performances comparisons on our method and three different methods.



(a) Retention Rate (b) Pseudo-label Accuracy (c) Clustering Accuracy

Figure 6: Effects of different pseudo-label selection thresholds on STL-10 and CIFAR-10.

Sensitivity to pseudo-label selection threshold. We further investigate the effect of the confidence threshold used for pseudo-label selection. As shown in Fig. 6, increasing the threshold improves pseudo-label accuracy but reduces sample retention, revealing a trade-off between label quality and sample coverage. Nevertheless, the final clustering performance remains relatively stable within the threshold range (0.7–1.0), supporting our choice of 1.0 as a unified default setting.

5 Conclusion

In this paper, we propose a novel Language-Assisted Image Clustering (LAIC) framework that more effectively exploits the semantic knowledge embedded in VLMs. By revisiting the roles of text modality in existing LAIC methods, we observe two key limitations of pair-centric paradigm, including the weak discriminability of constructed text features and the restricted use of manually defined semantic spaces during clustering. To overcome these issues, we introduce a cross-modal relation mining strategy that constructs an image–text representation matrix, enabling the extraction of more discriminative representations. Furthermore, we propose a category-wise prompt learning scheme that learns continuous semantic centers from the native semantic space of the VLMs. By optimizing these centers to align with image features, our method goes beyond fixed image-text alignments and achieves more accurate and flexible semantic representations for clustering. Experimental results show that our method outperforms state-of-the-art methods on eight benchmark datasets.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants U24A20322, 62576094 and 62422118, in part by the Hong Kong UGC under grants UGC/FDS11/E03/25 and UGC/FDS11/E04/26, and in part by the Hong Kong Research Grants Council under Grants RFS2627-1S02, R1018-25F. This research work was also supported by the Big Data Computing Center of Southeast University.

References

[1] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pages 268–285, 2020.

[2] Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7509–7524, 2023.

[3] Zhixin Li and Yuheng Jia. Conmix: Contrastive mixup at representation level for long-tailed deep clustering. In *International Conference on Learning Representations*, 2025.

[4] Zhengxing Jiao, Yaxin Hou, Jun Ma, Yuhang Li, Ding Ding, Yuheng Jia, Hui Liu, and Junhui Hou. Samples are not equal: A sample selection approach for deep clustering. In *International Conference on Learning Representations*, 2026.

[5] Zhixin Li, Yuheng Jia, Guanliang Chen, Hui Liu, and Junhui Hou. Mini-cluster guided long-tailed deep clustering. In *International Conference on Learning Representations*, 2026.

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

[7] Shaotian Cai, Liping Qiu, Xiaojun Chen, Qin Zhang, and Longteng Chen. Semantic-enhanced image clustering. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pages 6869–6878, 2023.

[8] Liping Qiu, Qin Zhang, Xiaojun Chen, and Shaotian Cai. Multi-level cross-modal alignment for image clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14695–14703, 2024.

[9] Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. In *International Conference on Machine Learning*, pages 27890–27902, 2024.

[10] Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. On the provable importance of gradients for autonomous language-assisted image clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19805–19815, 2025.

[11] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[12] James B McQueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[13] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[14] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pages 478–487, 2016.

[15] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, pages 132–149, 2018.

[16] Chuang Niu, Hongming Shan, and Ge Wang. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31:7264–7278, 2022.

[17] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8547–8555, 2021.

[18] Yunfan Li, Mouxing Yang, Dezhong Peng, Taihao Li, Jiantao Huang, and Xi Peng. Twin contrastive learning for online clustering. *International Journal of Computer Vision*, 130(9):2205–2221, 2022.

[19] Yuheng Jia, Jianhong Cheng, Hui Liu, and Junhui Hou. Towards calibrated deep clustering network. In *International Conference on Learning Representations*, 2025.

[20] Hanyang Li, Yuheng Jia, Hui Liu, and Junhui Hou. You can trust your clustering model: a parameter-free self-boosting plug-in for deep clustering. In *Advances in Neural Information Processing Systems*, 2025.

[21] Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models. In *British Machine Vision Conference*, 2023.

[22] Tianzhe Chu, Shengbang Tong, Tianjiao Ding, Xili Dai, Benjamin Haeffele, Rene Vidal, and Yi Ma. Image clustering via the principle of rate reduction in the age of pretrained models. In *International Conference on Learning Representations*, 2024.

[23] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

[24] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8150–8159, 2019.

[25] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8849–8858, 2020.

[26] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.

[27] Tsung Wei Tsai, Chongxuan Li, and Jun Zhu. Mice: Mixture of contrastive experts for unsupervised image clustering. In *International Conference on Learning Representations*, 2021.

[28] Huasong Zhong, Jianlong Wu, Chong Chen, Jianqiang Huang, Minghua Deng, Liqiang Nie, Zhouchen Lin, and Xian-Sheng Hua. Graph contrastive clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9224–9233, 2021.

[29] Yuming Shen, Ziyi Shen, Menghan Wang, Jie Qin, Philip Torr, and Ling Shao. You never cluster alone. In *Advances in Neural Information Processing Systems*, volume 34, pages 27734–27746, 2021.

[30] Hongyu Li, Lefei Zhang, and Kehua Su. Dual mutual information constraints for discriminative clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8571–8579, 2023.

[31] Zhixin Li, Yuheng Jia, Hui Liu, and Junhui Hou. Learning from sample stability for deep clustering. In *International Conference on Machine Learning*, 2025.

[32] Xianghan Meng, Zhiyuan Huang, Wei He, Xianbiao Qi, Rong Xiao, and Chun-Guang Li. Exploring a principled framework for deep subspace clustering. In *International Conference on Learning Representations*, 2025.

[33] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5879–5887, 2017.

[34] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011.

[35] Alex Krizhevsky. Learning multiple layers of features from tiny images. *https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf*, 2009.

[36] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.

[37] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[38] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[41] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900, 2022.

[42] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[43] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400, 2019.

[44] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

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

[46] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[47] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[48] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[49] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[50] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

Table of Contents for the Appendix

- A: Datasets Details 15
- B: Implementation Details 15
- C: More Experiment Results and Analyses 16
- C.1: Experimental Results under Multiple Random Seeds 16
- C.2: Experimental Results on Noisy and Weak Language Signals 16
- C.3: Experimental Results in More Complex Scenarios 17
- D: More Ablation Studies and Analyses 18
- D.1: Ablation Study of Semantic-center Parameterization 18
- D.2: Ablation Study of Pseudo-label Selection Strategy 19
- D.3: Ablation Study of Learning Independent Cluster Centroids 19
- E: Hyper-parameter Analyses 20
- F: Limitations 21
- G: Compute Resources 21
- H: Broader Impacts 21

A Datasets Details

We conduct experiments on eight datasets that cover a broad spectrum of clustering scenarios, ranging from general object recognition to fine-grained recognition, texture understanding, and action-oriented visual semantics, allowing us to comprehensively evaluate the effectiveness and generalizability of our method. Specifically, the datasets used in our experiments are described as follows:

- **ImageNet-Dogs** [33] is a fine-grained subset derived from ImageNet, where all categories correspond to different dog breeds. Since different dog breeds often share highly similar global shapes and visual appearances, this dataset requires the model to capture subtle discriminative cues.
- **ImageNet-10** [33] consists of ten semantically distinct categories selected from ImageNet. Compared with ImageNet-Dogs, this dataset focuses more on coarse-grained general object recognition, where categories usually exhibit larger inter-class differences.
- **CIFAR-10** [35] is a classic low-resolution image classification benchmark containing ten general object categories. Each image has a resolution of 32×32 pixels, making the visual details relatively limited.
- **STL-10** [34] contains ten common object categories, including animals and vehicles, with relatively higher-resolution images than CIFAR-style datasets.
- **Flowers-102** [36] is a fine-grained flower recognition dataset containing 102 categories. Many flower species have similar shapes and colors, while intra-class variations can also be large due to differences in viewpoint. It is challenging for clustering methods and requires discriminative semantic representations.
- **CIFAR-100** [35] extends the CIFAR-style benchmark to 100 categories. Compared with CIFAR-10, it contains a much larger number of fine-grained categories while keeping the same low image resolution.
- **DTD** [37] is a describable texture dataset containing 47 texture categories. Different from object-centric datasets, DTD focuses on material and texture patterns rather than object shapes. It is included to evaluate whether the proposed method can handle non-object-centric visual semantics.
- **UCF-101** [38] is a large-scale human action recognition dataset containing 101 action categories collected from realistic videos. The categories involve diverse human-object interactions, body motions, and scene contexts.

Overall, these datasets provide a comprehensive evaluation protocol from simple to complex scenarios and from general to specialized visual semantics. The number of samples and categories for each dataset is summarized in Table 9.

Table 9: A summary of benchmark datasets used for evaluation.

Dataset	#Samples	#Classes
ImageNet-Dogs [33]	19,500	15
ImageNet-10 [33]	13,000	10
STL-10 [34]	13,000	10
CIFAR-10 [35]	60,000	10
Flowers-102 [36]	8189	102
CIFAR-100 [35]	60,000	100
DTD [37]	5,640	47
UCF-101 [38]	13,320	101

B Implementation Details

Following previous works [7, 9], we adopt the pre-trained CLIP model with ViT-B/32 [39] and Transformer [40] as image and text backbones, respectively. We adopt data augmentation methods from [9] as weak augmentation $\mathcal{A}_w$, and incorporate RandAugment [44] to construct strong augmentation $\mathcal{A}_s$. Following [45], training is done for 20 epochs using batch size of 32, with SGD optimizer and an

initial learning rate of $2e-3$, which is decayed by the cosine annealing rule. We fix $\theta = 2, \gamma = 5$ and $\hat{k} = 10$ in all the experiments. The only exception is that we set $\hat{k} = 1$ on datasets with larger cluster numbers, including Flowers-102, CIFAR-100, DTD and UCF-101. All hyperparameters are selected without using the test set or any tuning procedure, and our method remains robust to their values within a reasonable range All experiments are conducted on a single NVIDIA RTX 4090 GPU.

We evaluate clustering performance using three widely adopted metrics: Normalized Mutual Information (NMI), Accuracy (ACC), and Adjusted Rand Index (ARI). NMI and ACC range in $[0, 1]$, while ARI ranges in $[-1, 1]$. For all three metrics, higher scores correspond to better clustering performance.

C More Experiment Results and Analyses

C.1 Experimental Results under Multiple Random Seeds

To further evaluate the stability and robustness of our method, we report the average experimental results and standard deviations of our method and the comparison methods over 5 runs with different random seeds. The experiments are conducted on all datasets described in Sec. 4.1, and the results are presented in Table 10 and Table 11. As can be observed, our method consistently achieves superior average performance across different datasets and evaluation metrics. More importantly, the standard deviations of our method are generally comparable to or smaller than those of existing methods, indicating that the performance gains are not caused by favorable random initialization or occasional runs. Instead, the proposed cross-modal relational supervision provides stable and reliable clustering guidance under different random seeds. These results further demonstrate that our method not only improves clustering accuracy, but also exhibits strong robustness and reproducibility across diverse datasets.

Table 10: Average clustering performance (in percent %) across four widely used benchmarks datasets.

Datasets	ImageNet-Dogs			ImageNet-10			STL-10			CIFAR-10		
Methods	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
SIC [7]	69.0±1.6	69.7±1.1	55.8±1.5	96.9±0.2	98.3±0.1	96.2±0.2	95.3±0.1	98.1±0.1	95.9±0.1	84.7±0.1	92.6±0.1	84.4±0.1
MCA [8]	73.3±1.5	74.9±2.5	61.6±2.5	-	-	-	95.5±0.1	98.1±0.1	96.0±0.1	84.9±0.2	92.7±0.2	84.6±0.2
TAC [9]	80.9±1.0	83.3±1.4	72.0±1.0	98.2±0.1	99.2±0.1	98.3±0.1	95.6±0.2	98.3±0.1	96.1±0.1	83.4±0.2	92.1±0.2	83.4±0.3
LFSS [31]	61.9±1.9	69.3±1.7	53.1±1.6	85.2±0.6	93.8±0.5	85.4±0.8	77.3±1.1	86.1±0.9	74.0±0.8	84.4±0.6	92.2±0.4	84.6±0.5
PRO-DSC [32]	-	-	-	98.0±0.3	99.0±0.2	97.8±0.2	95.4±0.2	98.1±0.4	95.9±0.2	79.6±0.4	87.1±0.3	80.2±0.4
Ours	**86.2**±1.1	**88.0**±1.2	**82.6**±1.2	**99.6**±0.1	**99.8**±0.1	**99.7**±0.1	**96.1**±0.2	**98.5**±0.1	**96.7**±0.1	**85.2**±0.3	**92.9**±0.3	**84.8**±0.2

Table 11: Average clustering performance (in percent %) across four benchmarks datasets with larger cluster numbers.

Datasets	Flowers-102			CIFAR-100			DTD			UCF-101		
Methods	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
SIC [7]	86.9±0.9	70.2±0.8	64.2±0.6	63.5±0.8	48.3±0.7	34.7±0.9	59.9±0.8	45.5±0.8	30.8±0.9	81.7±0.8	62.1±1.1	52.5±0.9
TAC [9]	87.7±0.8	70.8±0.9	63.9±0.5	68.1±0.2	55.4±0.7	42.4±0.6	61.8±0.5	50.9±0.6	34.8±0.7	82.5±0.4	69.1±0.8	60.2±0.7
LFSS [31]	79.2±1.1	55.4±1.2	51.8±0.9	66.0±0.6	52.4±0.8	37.6±0.8	58.7±1.1	45.5±1.2	31.9±0.9	58.9±0.9	34.5±1.2	21.6±1.1
PRO-DSC [32]	83.2±1.0	69.8±0.8	59.6±0.7	68.2±0.5	55.7±0.6	40.2±0.3	57.6±0.7	48.1±0.9	31.3±0.8	81.0±0.8	61.9±0.9	52.8±0.7
Ours	**89.6**±0.8	**78.6**±0.7	**72.9**±0.4	**68.8**±0.2	**58.1**±0.4	**43.4**±0.3	**63.8**±0.6	**53.7**±0.6	**36.2**±0.5	**84.4**±0.5	**70.2**±0.7	**61.9**±0.9

C.2 Experimental Results on Noisy and Weak Language Signals

In practical scenarios, language knowledge extracted may contain noise or may be incomplete. Therefore, we conduct two additional experiments to analyze whether our method can still extract reliable signals under such conditions.

Robustness to noisy language signals. To evaluate robustness to noisy language signals, we perturb the features extracted from WordNet. Specifically, a certain ratio of the language features is randomly selected and corrupted by adding Gaussian noise. We then evaluate the performance under different noise ratios. The results are shown in Table 12. Compared with TAC, our method consistently exhibits smaller performance degradation as the noise ratio increases. For example, under 30% noise,

Table 12: Robustness to noisy language signals. We report the average score and the performance decline compared with the noise-free setting.

Dataset	Noise Ratio	Ours		TAC [9]	
		Score Avg.	Decline	Score Avg.	Decline
CIFAR10	0%	87.6	0.0	86.1	0.0
	10%	87.4	**0.2**	85.7	0.4
	30%	84.9	**2.7**	83.0	3.1
DTD	0%	51.2	0.0	48.8	0.0
	10%	50.9	**0.3**	48.2	0.6
	30%	47.2	**4.0**	43.1	5.7

Table 13: Robustness to weak language signals. We randomly retain different ratios of WordNet to simulate weak language knowledge.

Dataset	WordNet Ratio	Ours		TAC [9]	
		Score Avg.	Decline	Score Avg.	Decline
CIFAR10	100%	87.6	0.0	86.1	0.0
	10%	86.8	**0.8**	84.9	1.2
	1%	85.1	**2.5**	79.4	6.7
DTD	100%	51.2	0.0	48.8	0.0
	10%	50.1	**1.1**	45.1	3.7
	1%	46.3	**4.9**	39.6	9.2

our method drops by 2.7% and 4.0% on CIFAR10 and DTD, respectively, while TAC drops by 3.1% and 5.7%. This suggests that our method is more robust to noisy language signals. One possible reason is that the second stage of our method does not directly rely on the textual features constructed in the first stage, whereas TAC still depends on these textual features during training.

Robustness to weak language signals. We further evaluate the robustness of our method when only weak language signals are available. Specifically, we randomly retain a subset of WordNet with different ratios to simulate incomplete language knowledge. A smaller WordNet ratio indicates weaker language signals. The results are reported in Table 13. Our method again shows stronger robustness than TAC. When only 1% of WordNet is retained, our method suffers a decline of 2.5% points on CIFAR10 and 4.9% on DTD, while TAC drops by 6.7% and 9.2%, respectively.

The reason is that when the available corpus becomes smaller, text-based methods relying on neighbor matching may struggle to find appropriate nouns. In contrast, even with a reduced corpus, the remaining words can still serve as semantic anchors in our framework. By measuring the cross-modal relationships between samples and these semantic anchors, our method can still extract useful supervisory information from weak language signals.

C.3 Experimental Results in More Complex Scenarios

To further evaluate the effectiveness of LAIC methods in more complex scenarios, we conduct additional experiments on four challenging datasets involving fine-grained categories, remote-sensing scenes, multi-object contexts, and compositional attributes.

Specifically, we include two existing datasets: **Stanford Cars** [46], a fine-grained vehicle recognition dataset with subtle inter-class differences, and **EuroSAT** [47], a satellite image dataset whose categories are defined by remote-sensing scenes rather than common object concepts. In addition, we construct two datasets from public sources to evaluate more complex semantic structures. First, we construct **Scene-6** (shown in Fig. 7) from the multi-label datasets COCO [48] and Visual Genome [49], both of which contain real-world complex scenes. We select six scene categories, such as people eating and people exercising on grass, with 100 images per category. These categories usually involve multiple objects, human-object interactions, and contextual scene information, and therefore cannot be directly represented by a single WordNet word. Second, we construct **Face-4** (shown in Fig. 8) from CelebA [50] by selecting four compositional categories: men with glasses, women with glasses,

men without glasses, and women without glasses, with 2,000 images per category. This dataset evaluates whether the clustering method can handle attribute compositions involving both gender and glasses.



Bathroom | Vehicles on road | People eating | People on grass | People using PC | People near water

Figure 7: Example images from **Scene-6** (constructed by us).



A man with glasses | A man without glasses | A woman with glasses | A woman without glasses

Figure 8: Example images from **Face-4** (constructed by us).

Table 14: Clustering performance (in percent %) in more complex scenarios.

Datasets	Stanford Cars			EuroSAT			Scene-6			Face-4			Avg.
Methods	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
SIC [7]	58.8	22.4	15.9	54.3	63.8	44.9	89.7	93.9	88.1	42.1	51.6	28.5	54.5
CPP [22]	67.8	32.4	26.9	62.3	70.8	52.9	90.7	94.9	89.1	46.1	61.6	31.5	60.6
TAC [9]	60.2	24.8	17.6	57.9	66.9	48.7	91.0	95.0	90.3	43.7	53.9	30.9	56.7
Ours	**73.3**	**48.1**	**36.4**	**65.4**	**74.4**	**56.0**	**96.2**	**98.3**	**95.9**	**51.2**	**64.8**	**41.3**	**66.8**

We compare our method with representative LAIC methods, including SIC [7] and TAC [9], as well as CPP [22] (a deep clustering method that uses only CLIP image features). The experimental results in Table 14 show that our method consistently outperforms these baselines on all four complex datasets. These results demonstrate that our method is not limited to simple object-centric datasets or categories that can be directly described by predefined words. By considering all candidate texts, our method can associate images with richer semantic information, enabling more comprehensive semantic capture for complex images. It can more effectively exploit textual semantics in complex scenarios involving fine-grained categories, remote-sensing scenes, multi-object interactions, and compositional attributes.

D More Ablation Studies and Analyses

D.1 Ablation Study of Semantic-center Parameterization

We include anotheclustering method, denoted as Ours (MLP-head). Specifically, using the high-quality pseudo-labels selected in Sec. 3.1 together with image features, we train a clustering head with the loss in Eq. (11). The clustering head is implemented as a two-layer MLP with dimensions 512-512-K.

The results are shown in Table 17. Directly using the first-stage learned matrix $\mathbf{C}$ already provides competitive clustering performance. Our final design, which uses the learned semantic category centers as clustering anchors, achieves the best performance across all datasets and metrics. This

demonstrates that the learned semantic centers provide more effective and semantically meaningful anchors for clustering.

Table 15: Ablation study of learning independent cluster centroids. Original denotes our default method, where semantic centers are parameterized through the frozen CLIP text encoder. Independent centers denotes directly learning cluster centroid variables without using the text encoder.

Datasets	CIFAR-10			DTD-47			UCF-101		
Methods	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
Original	**85.2**	**92.9**	**84.8**	**63.8**	**53.7**	**36.2**	**76.1**	**60.9**	**51.4**
Independent centers	82.7	88.6	81.6	62.1	51.4	33.4	74.6	57.3	48.6

D.2 Ablation Study of Pseudo-label Selection Strategy

We further compare our pseudo-label selection strategy with confidence-based selection methods. Specifically, we apply the softmax operation to the cosine similarities between samples and clustering centers, and use the probability assigned to the corresponding center as the confidence score. Based on this confidence score, we consider two types of high-confidence sample selection strategies: Global_x, which selects the top confident samples with ratio x globally, and Class_x, which selects the top confident samples with ratio x within each class. We compare these variants with our default method, denoted as Original.

As shown in Table 16, confidence-based selection generally leads to inferior performance compared with our method. Our nearest-neighbor pseudo-label consistency strategy additionally considers feature-level consistency. In the space of $\mathbf{C}$, a sample and its nearest neighbors have a high probability of sharing the same ground-truth label. Therefore, nearest-neighbor consistency provides a more reliable criterion for selecting high-quality pseudo-labels.

Table 16: Ablation study of pseudo-label selection strategies. Global_x selects the top confident samples with ratio x globally, while Class_x selects the top confident samples with ratio x from each class. Original denotes our default nearest-neighbor consistency based selection strategy.

Datasets	CIFAR-10			DTD		
Methods	NMI	ACC	ARI	NMI	ACC	ARI
Original	**85.2**	**92.9**	**84.8**	**63.8**	**53.7**	**36.2**
Global_0.9	84.2	91.4	83.7	62.7	52.0	34.7
Global_0.8	84.3	91.8	84.0	62.9	52.2	35.2
Global_0.7	83.8	92.0	83.4	61.9	52.6	34.8
Class_0.9	83.7	91.6	82.6	62.4	51.4	34.7
Class_0.8	84.4	92.1	83.4	62.8	52.1	34.9
Class_0.7	83.9	91.7	82.9	61.6	51.9	33.9

D.3 Ablation Study of Learning Independent Cluster Centroids

We also study whether the semantic centers should be parameterized through the CLIP text encoder. To this end, we construct a variant that independently learns a set of cluster centroid variables without using the text encoder. This variant directly uses the original pseudo-labels and the same training loss to optimize the centers, denoted as Independent centers.

The results are reported in Table 17. Compared with our original design, independently learning cluster centroids leads to consistent performance degradation on all datasets. We attribute this degradation to the fact that independently learned centers are not constrained to lie in the native semantic space of CLIP, which may weaken their generalization ability and semantic expressiveness. In contrast, our method learns semantic centers through a frozen CLIP text encoder, which imposes a structural constraint on center learning and ensures that the learned centers remain aligned with the semantic space of CLIP.

Table 17: Ablation study of learning independent cluster centroids. Original denotes our default method, where semantic centers are parameterized through the frozen CLIP text encoder. Independent centers denotes directly learning cluster centroid variables without using the text encoder.

Datasets	CIFAR-10			DTD-47			UCF-101		
Methods	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
Original	**85.2**	**92.9**	**84.8**	**63.8**	**53.7**	**36.2**	**76.1**	**60.9**	**51.4**
Independent centers	82.7	88.6	81.6	62.1	51.4	33.4	74.6	57.3	48.6

E Hyper-parameter Analyses

Number of closed nouns selected for fine-grained image center θ. In Sec. 3.1, we first obtain fine-grained image centers and assign θ nouns to each center to represent its semantics. To investigate the effect of θ, we evaluate the performance of directly applying K-means on $\mathbf{C}$ (noted as "no-train") and the final clustering performance under different values of θ on CIFAR-10 and DTD. We observe that when θ is small, the clustering performance remains relatively stable, while it degrades noticeably as θ increases. We think that a small θ allows the selected candidate nouns to accurately capture the semantic of different classes. In this case, nouns that are close to a given image provide reliable semantic descriptions. Meanwhile, distant nouns are likely associated with other classes. They can provide meaningful discriminative information, indicating that the image is not close to a certain other class. However, when θ becomes large, the candidate noun set tends to include more nouns that are irrelevant to the dataset, which affects the quality of the $\mathbf{C}$. Based on these observations, we set $\theta = 2$ in all experiments.

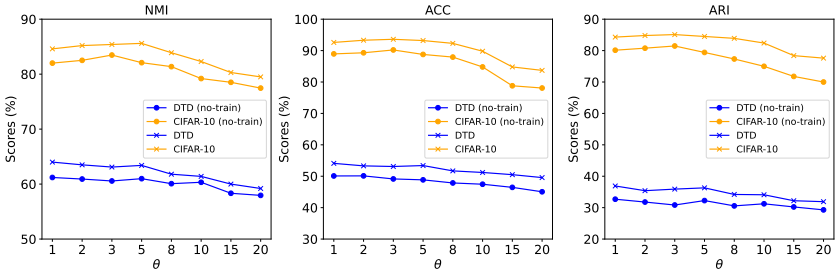


Figure 9: Clustering performances with different θ on CIFAR-10 and DTD.

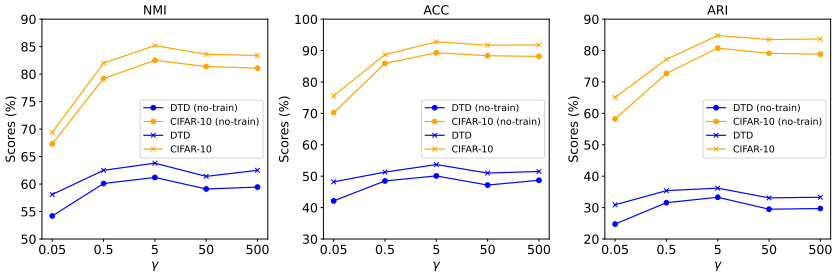


Figure 10: Clustering performances with different γ on CIFAR-10 and DTD.

Regularization parameter γ. In Problem (3), we use Frobenius regularizer to impose a smoothness constraint on $\mathbf{C}$, preventing it from being overly sparse, and γ is the regularization parameter. We obtain the matrix $\mathbf{C}$ under different values of γ on CIFAR-10 and DTD, and report both the clustering results obtained by directly applying K-means on $\mathbf{C}$ (noted as "no-train") and the final clustering performance (shown in Fig. 10). We observe that the clustering performance degrades noticeably when γ is smaller than 5, while it remains relatively stable as γ is set larger. When γ is overly small, the matrix becomes much sparse, which limits the number of text features involved in describing the image, resulting in worse discriminability. This suggests that describing the image from the global candidate noun set is appropriate. When γ is too large, the matrix becomes overly smooth,

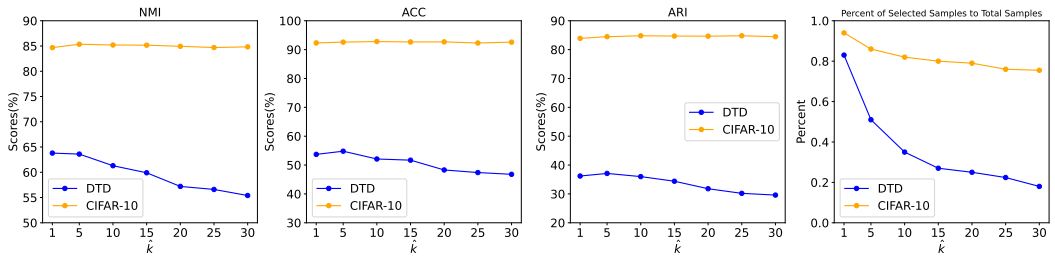


Figure 11: Clustering performances and percent of selected samples to total samples with different $\hat{k}$ on CIFAR-10 and DTD.

with each row becoming more similar and introducing redundant information, and also reduce the discriminability of $\mathbf{C}$. Our method is robust to the choice of γ within a reasonably large range, and is primarily sensitive only to overly small values of the regularization parameter. We fix $\gamma = 5$ in all experiments to achieve a favorable balance between reconstruction accuracy and richness.

Number of nearest neighbors $\hat{k}$. In Sec. 3.2, we filter high-quality samples by measuring the consistency between each sample and its $\hat{k}$ nearest neighbors. We observe that even with a large $\hat{k}$, a sufficient number of high-quality samples can still be selected on CIFAR-10, making the final clustering performance relatively insensitive to the choice of $\hat{k}$. In contrast, the DTD dataset has a much larger clustering number of. As $\hat{k}$ increases, the requirement that a sample and all its $\hat{k}$ nearest neighbors share the same pseudo-label becomes overly strict. This leads to an obvious reduction in the number of selected samples and degrades clustering performance. Thus, for datasets with a smaller number of clusters, the choice of $\hat{k}$ can be relatively flexible, which we set to 10. For datasets with a larger number of clusters, we set $\hat{k} = 1$ to balance both the quantity and quality of the selected samples.

F Limitations

Although the proposed cross-modal relation mining method can effectively extract more discriminative self-supervised signals, further improving the utilization efficiency of the textual modality remains a challenging task. In future work, we aim to further refine this framework to better address this issue.

G Compute Resources

- CPU: AMD EPYC 7642 48-Core Processor × 2
- GPU: NVIDIA GeForce RTX 4090 24G × 1
- MEM: 500G
- Maximum total computing time: training + testing within 26min

H Broader Impacts

This paper presents work whose goal is to advance the field of Machine Learning. Specifically, we propose a novel Language-Assisted Image Clustering framework that more effectively exploits the semantic knowledge embedded in vision-language models for unsupervised clustering. By mining cross-modal relational signals and learning continuous category-wise semantic centers from the native semantic space of VLMs, our method obtains more discriminative representations, improves clustering performance without additional supervision, and provides interpretable semantic descriptions for the discovered clusters. Experimental results on eight benchmark datasets show that our method outperforms state-of-the-art approaches, highlighting the potential of better utilizing both image and text modalities for clustering tasks. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope. Please see Abstract and Sec. 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work performed by the authors. Please see Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not). Please see Sec. 4.1 and Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results. Please see Sec. 4.1 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports mean and standard deviation over multiple random seeds for the main experiments in Appendix C.1, which provides information about the stability and statistical reliability of the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments. Please see Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both potential positive societal impacts and negative societal impacts of the work performed. Please see Appendix H.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our research does not release data or models, so the paper has no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, are properly credited and are the license and terms of use explicitly mentioned and properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.