FROM LEARNING TO UTILIZATION: BETTER EXPLOITING EXTERNAL KNOWLEDGE FOR IMAGE CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised image clustering increasingly leverages external knowledge, including textual vocabularies, LLM-generated descriptions, and spatial priors. However, existing studies mainly focus on how to learn useful external knowledge into image representations, while overlooking whether the learned knowledge is effectively utilized in the final cluster assignments. In this work, we identify a previously underexplored *learning-utilization gap*: useful category-related knowledge may be encoded in the representation but remain underutilized in clustering predictions. We formalize learning failure, utilization failure, and their gap from an information-theoretic perspective, and further show that distance-based structural relations with correctly predicted neighbors can recover underutilized knowledge and correct erroneous predictions. Guided by this theory, we propose CARE. CARE first establishes stable initial cluster assignments from externally enhanced representations through iterative pseudo-label optimization, then evaluates sample-level knowledge-utilization reliability, and finally preserves reliable predictions while adaptively exploiting local feature-space structure to refine unreliable ones. In this way, CARE progressively better aligns cluster assignments with the category-related knowledge encoded in the learned representation. Extensive experiments on 6 widely used benchmark datasets and 3 more challenging datasets, covering 3 different types of external auxiliary information, demonstrate that CARE improves clustering assignments. The code will be published.

1 INTRODUCTION

Image clustering aims to group semantically homogeneous samples into the same clusters from unlabeled images (Van Gansbeke et al., 2020). Recently, an increasing number of image clustering methods utilize external knowledge, such as predefined vocabularies (Li et al., 2023; Zhu et al., 2026), LLM-generated descriptions (Li et al., 2026; Zhong et al., 2026), and spatial priors (Jiu et al., 2026), to improve image features and facilitate image clustering. Despite this progress, most studies primarily focus on how to **learn** useful knowledge, while paying much less attention to whether the learned knowledge is effectively **utilized** in the final cluster assignments.

A related distinction has long been recognized in cognitive psychology. Tulving and Pearlstone (Tulving & Pearlstone, 1966) distinguished *availability* from *accessibility*, showing that information may remain available in memory while being inaccessible during re-

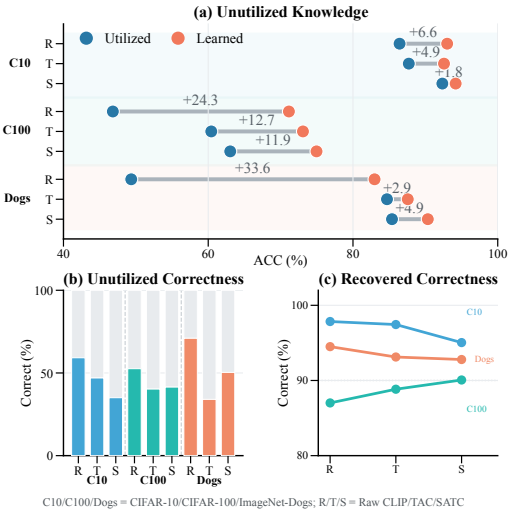


Figure 1: **Our observations.** Learned knowledge is not fully utilized in final cluster assignments. Many misclustered samples still retain class-consistent structure, which constitutes the primary source of successful CARE corrections.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

trieval. This distinction was further developed through the notions of *storage strength* and *retrieval strength* (Bjork et al., 1992), and more broadly through the distinction between *learning* and *performance* (Soderstrom & Bjork, 2015). Inspired by this perspective, we ask whether an analogous phenomenon exists in external-knowledge-assisted image clustering: *can useful knowledge be encoded in an image representation, yet remain underutilized in the final cluster assignment?*

Our observations suggest that the answer is yes. As shown in Fig. 1(a), we examine three frozen representations on CIFAR-10, CIFAR-100, and ImageNet-Dogs: raw CLIP features, TAC features enhanced by textual knowledge, and SATC features enhanced by selectively utilized textual and spatial priors. For each representation, we compare the final cluster assignments obtained via the readout in Sec. 4.2 with an analysis-only kNN oracle(see Appendix A.2 for detailed settings), revealing a substantial gap between what has been learned in the frozen representation and what is actually utilized in the final assignments. Compared with raw CLIP, TAC and SATC markedly reduce this gap, indicating that external knowledge enriches the representation structure but does not eliminate the utilization gap. We further analyze the misclustered samples on the three datasets. As shown in Fig. 1(b), a substantial proportion still retain correct class-consistent structure in the frozen representation, revealing that learned structure does not always translate into correct predictions. We therefore distinguish two sources of clustering error: **learning failure**, where useful category-related structure has not been learned, and **utilization failure**, where such structure is already present but not sufficiently reflected in the final assignment. This empirical distinction leads to a central question: *how can we better exploit external knowledge to reduce utilization failure?*

To address these challenges, we first formalize *learning failure*, *utilization failure*, and the *learning-utilization gap* from an information-theoretic perspective, theoretically establishing that useful knowledge can be successfully learned into an image representation while still failing to be adequately reflected in its final cluster assignment. We further show that informative local neighborhood consistency in the frozen image representation can be effectively exploited to refine cluster assignments and recover underutilized knowledge. Specifically, when a misclustered sample is sufficiently close in feature space to correctly predicted neighbors that collectively provide stronger support for its true category than for the erroneous one, incorporating these distance-based structural relations can reverse the prediction error.

Guided by this theory, we propose CARE, an effective framework that better translates learned knowledge into cluster assignments through three stages: Assignment Establishment, Reliability Assessment, and Knowledge Utilization. CARE first iteratively optimizes pseudo-labels on the frozen knowledge-enhanced representation to obtain stable and reliable initial cluster assignments. CARE then assesses the sample-level reliability of these stabilized predictions according to their local structural consistency. Finally, it preserves highly reliable assignments while adaptively introducing informative neighborhood structure for less reliable samples, thereby recovering residual category-related knowledge that remains encoded in the representation but is still insufficiently reflected in the cluster assignments. Moreover, we further analyze the samples correctly recovered by CARE. As shown in Fig. 1(c), more than 88% of these recovered samples exhibit strong class-consistent neighborhood structure in a post-hoc ground-truth analysis, providing substantial empirical support for our theoretical conclusion that distance-based structural relations can effectively recover underutilized knowledge and correct erroneous predictions.

The main contributions of the paper are as follows:

- We identify the learning-utilization gap in external-knowledge-assisted image clustering and provide an information-theoretic characterization of learning failure, utilization failure, and their gap. We further theoretically show that distance-based structural relations with correctly predicted neighbors can reverse the prediction error of misclustered samples.
- We propose a theory-guided framework CARE, which establishes stable initial cluster assignments, assesses sample-level knowledge-utilization reliability, and adaptively exploits local feature-space structure to refine unreliable predictions, thereby recovering knowledge that has already been learned but remains underutilized.
- We conduct extensive experiments on 6 widely used benchmark datasets and 3 more challenging datasets with 3 different types of external auxiliary information, demonstrating that CARE consistently improves clustering performance and validating the effectiveness of better utilizing learned external knowledge.

2 Related Work

2.1 External Knowledge-Assisted Image Clustering

Image clustering aims to partition unlabeled images into semantically meaningful groups without access to ground-truth annotations (Van Gansbeke et al., 2020). Early methods mainly construct internal supervision signals from the images themselves, such as pseudo-labels, neighborhood consistency, and augmentation invariance. With the emergence of large-scale pretrained vision-language models, recent methods increasingly exploit external semantic knowledge. SIC Cai et al. (2023) maps images into a CLIP-guided semantic space for consistency learning. TAC Li et al. (2023) introduces predefined textual concepts and transfers their semantic knowledge into image representations through cross-modal learning. CAE Zhu et al. (2026) integrates noun-level concepts with caption-level descriptions to capture complementary high-level and fine-grained semantics for image clustering. GSEC Li et al. (2026) leverages LLM-generated descriptions to provide richer and more adaptive textual knowledge for image clustering. KEC Zhong et al. (2026) constructs hierarchical concept-attribute knowledge with LLMs to provide fine-grained and discriminative semantic guidance for image clustering. SATC Jiu et al. (2026) combines selectively utilized textual knowledge with spatial structural priors to enrich useful information in image representations. Despite these advances, existing methods primarily focus on acquiring, selecting, or transferring external knowledge to improve visual representations or clustering predictions. Less attention has been paid to whether the knowledge encoded in the resulting representations is effectively exploited by the final cluster assignments.

2.2 Local Structure-based Learning and Refinement

Local structure has been widely exploited to improve prediction reliability by assuming that nearby samples in a meaningful feature space tend to share semantic labels. Zhou et al. (Zhou et al., 2003) propagate label information according to the geometry of labeled and unlabeled samples. LaplacianShot (Ziko et al., 2020) performs transductive inference on a fixed pretrained embedding by encouraging nearby query samples to have consistent assignments through Laplacian regularization. In image clustering, SCAN (Van Gansbeke et al., 2020) exploits neighborhood consistency to learn semantic clusters, while ConAff (Yu et al., 2023) mines contextually affinitive neighbors and suppresses noisy relations near cluster boundaries. More recently, DCBoost (Li et al., 2025a) selects reliable samples through adaptive kNN consistency filtering and uses them as anchors to further optimize the global feature structure. These studies demonstrate the effectiveness of local geometry for clustering and prediction refinement.

However, existing methods do not explicitly address the gap between knowledge encoded in representations and that utilized by final predictions. CARE addresses this gap by theoretically characterizing when distance-based local relations can reverse erroneous assignments and using them to recover underutilized knowledge from frozen representations.

3 Theorem of Knowledge Utilization

We provide a theoretical analysis of knowledge utilization in image clustering. Let $Y \in [K]$ denote the latent semantic category, used only for theoretical analysis, Z the learned image representation, and $P = h(Z)$ the probabilistic cluster assignment produced from Z. Our analysis characterizes (i) whether useful knowledge can be learned into Z but remain insufficiently reflected in P, and (ii) when such underutilized knowledge can be recovered through local feature geometry. All proofs are deferred to Appendix C. We build on standard identities of conditional entropy and conditional mutual information (Cover et al., 1991).

Theorem 1 (Learning-utilization decomposition) *Since P is obtained from Z, the residual semantic uncertainty satisfies*

$$H(Y \mid P) = H(Y \mid Z) + I(Y; Z \mid P). \tag{1}$$

Accordingly, $H(Y \mid Z)$ characterizes *learning failure*, i.e., semantic uncertainty that remains unresolved by the learned representation, whereas $H(Y \mid P)$ characterizes the remaining uncertainty at

the utilization stage. Their difference,

$$\mathcal{G}_{\text{util}} := H(Y \mid P) - H(Y \mid Z) = I(Y; Z \mid P), \tag{2}$$

defines the *learning-utilization gap*, quantifying category-related knowledge that has already been encoded in Z but is not reflected in P. Theorem 1 therefore formally confirms that successful knowledge learning does not necessarily imply complete knowledge utilization.

Definition 1 (Distance-based structural relation) *Let z_i and z_j be ℓ_2-normalized image representations, with cosine distance $d_{ij} = 1 - z_i^\top z_j$. For analysis only, let $\mathcal{A}^+ = \{j : \hat{y}_j = y_j\}$ denote correctly predicted samples after aligning cluster indices with categories, and let $\mathcal{N}_i^+$ contain the k nearest samples to i in $\mathcal{A}^+$, excluding i. The structural prediction is*

$$q_i = \frac{1}{|\mathcal{N}_i^+|} \sum_{j \in \mathcal{N}_i^+} p_j. \tag{3}$$

To study whether neighborhood predictions can correct an erroneous assignment, we refine the original prediction p_i toward the neighbor-aggregated prediction q_i with $\alpha_i \in [0, 1]$:

$$\widetilde{p}_i = p_i + \alpha_i(q_i - p_i) = (1 - \alpha_i)p_i + \alpha_i q_i, \quad \alpha_i \in [0, 1]. \tag{4}$$

Theorem 2 (Distance-based correction condition) *For a misclustered sample i with true category y_i, define $\delta_{ic} = p_{ic} - p_{iy_i}$ and $\gamma_{ic} = q_{iy_i} - q_{ic}$ for each $c \neq y_i$. If $\gamma_{ic} > 0$ for all $c \neq y_i$, then y_i becomes the unique argmax of $\widetilde{p}_i$ whenever*

$$\alpha_i > \max_{c \neq y_i} \frac{[\delta_{ic}]_+}{[\delta_{ic}]_+ + \gamma_{ic}}, \qquad [x]_+ = \max(x, 0). \tag{5}$$

Thus, local structure can correct an error when its weighted evidence outweighs the original prediction's erroneous preference. Define the class-consistent structural support as

$$\rho_i = \frac{|\{j \in \mathcal{N}_i^+ : y_j = y_i\}|}{|\mathcal{N}_i^+|}. \tag{6}$$

Corollary 1 (Structural-support monotonicity) *For each $c \neq y_i$, suppose the neighbor margin $p_{jy_i} - p_{jc}$ is at least $\mu_c^+ > 0$ for class-consistent neighbors and $-\mu_c^-$ for other neighbors, where $\mu_c^- \geq 0$. Then*

$$\gamma_{ic} \geq (\mu_c^+ + \mu_c^-)\rho_i - \mu_c^-. \tag{7}$$

With p_i, α_i, and the margin bounds fixed, increasing ρ_i raises the guaranteed refined margin for the true category, strengthening the sufficient guarantee for correction.

Together, these results motivate CARE: local structure can help recover category-related knowledge underutilized by the initial assignment, with stronger class-consistent support improving the correction guarantee under the stated assumptions.

4 Proposed Method

In this section, we describe the proposed CARE framework, illustrated in Fig. 2. CARE aims to better utilize external knowledge already encoded in image representations but underutilized by final cluster assignments. The overall pipeline consists of 4 stages: (a) External Knowledge Learning, (b) Assignment Establishment, (c) Reliability Assessment, and (d) Knowledge Utilization.

4.1 External Knowledge Learning

For the raw image dataset $\mathcal{D} = \{x_n\}_{n=1}^N$, we first use the pretrained CLIP image encoder to extract visual features $z_n^{\text{visual}} \in \mathbb{R}^d$, where d denotes the feature dimension. We introduce an auxiliary modality with features $z_n^{\text{aux}} \in \mathbb{R}^d$ to obtain knowledge-enhanced representations, constructed from textual vocabularies, LLM-generated descriptions, or spatial priors in our experiments; details are provided in Appendix A.2.

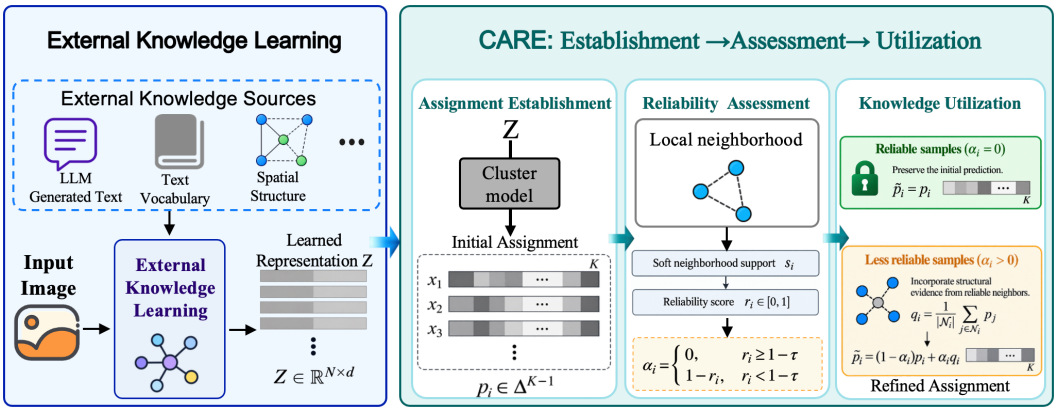


Figure 2: Overview of CARE. CARE establishes initial cluster assignments from image features already learned with external knowledge, assesses sample reliability from local structure, and adaptively refines unreliable predictions using local neighbors.

Let $\mathcal{M} = \{\text{visual}, \text{aux}\}$ denote the two involved modalities, and let c_i^m be the soft cluster assignment of sample i produced by the modality-specific clustering head for modality $m \in \mathcal{M}$. Following the original knowledge-learning pipelines (Li et al., 2023), external knowledge is transferred to the visual branch through mutual distillation:

$$\mathcal{L}_{\text{EK}} = \sum_{(m,n)\in\mathcal{E}} \mathcal{L}_{\text{distill}}^{m,n} + \lambda_1 \sum_{(m,n)\in\mathcal{E}} \mathcal{L}_{\text{consist}}^{m,n} - \lambda_2 \sum_{m\in\mathcal{M}} \mathcal{L}_{\text{entropy}}^{m}, \tag{8}$$

where $\mathcal{E}$ contains the modality pairs involved in mutual distillation. The three terms correspond to cross-modal knowledge transfer, prediction consistency, and entropy regularization, respectively.

After mutual distillation, we retain the knowledge-enhanced visual representations $Z = \{z_i\}_{i=1}^N$, where z_i is the hidden representation produced by the distilled visual clustering head, immediately before the final linear assignment layer. These features are subsequently frozen and used as the input to CARE. Although the formulation above uses one auxiliary modality for clarity, CARE operates on the resulting knowledge-enhanced visual representation and thus naturally extends to multiple auxiliary modalities, as further analyzed in Appendix H.

4.2 Assignment Establishment

After external knowledge learning, the frozen representation may already encode substantial category-related information, but converting it into reliable cluster assignments remains non-trivial. We therefore regard assignment establishment as the first stage of knowledge utilization.

Inspired by HUME (Gadetsky & Brbic, 2023), which suggests that semantically meaningful labelings should be linearly recoverable from informative representations, we introduce a linear inner learner over the frozen knowledge-enhanced representation. Its role is to provide a representation-supported readout that exposes information in the frozen representation that can be recovered by a linear predictor and incorporates it into the cluster assignments. Specifically, we freeze $Z = \{z_i\}_{i=1}^N$. At iteration t, the clustering model produces soft assignments $\tau_i^{(t)} \in \Delta^{K-1}$. A freshly initialized linear inner learner is trained for $S = 10$ gradient steps to minimize

$$\mathcal{L}_{\text{inner}}^{(t)}(w) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} \text{sg}[\tau_{ic}^{(t)}] \log\big[\text{softmax}(wz_i)\big]_c, \tag{9}$$

where $w \in \mathbb{R}^{K\times d}$ denotes the parameters of the linear readout and $\text{sg}[\cdot]$ denotes stop-gradient. Let $w^{(t,S)}$ denote the resulting parameters and $q_i^{(t)} = \text{softmax}(w^{(t,S)}z_i)$. Holding these predictions fixed, we update the clustering model using

$$\mathcal{L}_{\text{assign}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} \tau_{ic}^{(t)} \log \text{sg}[q_{ic}^{(t)}] - \gamma H\left(\frac{1}{N}\sum_{i=1}^{N}\tau_i^{(t)}\right), \tag{10}$$

where gradients flow through $\tau_i^{(t)}$ only, and the entropy term encourages balanced cluster usage.

The procedure is repeated, allowing the clustering assignments to progressively incorporate information recoverable from the frozen representation. The final soft prediction is denoted as $p_i = \tau_i \in \Delta^{K-1}$ for subsequent reliability assessment and knowledge utilization. We further analyze assignment iterations and alternative readout architectures in Appendix D and Appendix G.

4.3 Reliability Assessment

Theorem 2 motivates using reliable neighboring predictions to correct unreliable samples. Since ground-truth labels are unavailable, CARE estimates reliability from local prediction consistency on the frozen knowledge-enhanced representation Z. Specifically, for each sample i, let $\mathcal{N}_i$ denote its k nearest neighbors in a cosine kNN graph, and define

$$\bar{p}_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} p_j, \qquad s_i = p_i^\top \bar{p}_i, \tag{11}$$

where s_i measures the agreement between the sample prediction and its local neighborhood. To reduce sensitivity to the absolute scale of s_i across datasets, we use its empirical percentile rank as the reliability score and select samples in the top-τ percentile as reliable anchors:

$$r_i = \frac{\mathrm{rank}(s_i) - 1}{N - 1}, \qquad r_i \in [0, 1], \qquad \mathcal{A} = \{i \mid r_i \geq 1 - \tau\}. \tag{12}$$

Here, $\mathrm{rank}(s_i)$ denotes the ascending rank of s_i and $\mathcal{A}$ provides an observable proxy for the reliable samples in Theorem 2. The reliability of the selected anchors is further analyzed in Appendix E.

4.4 Knowledge Utilization

CARE preserves reliable anchors while using local structure to refine less reliable samples. For each $i \notin \mathcal{A}$, we retain the reliable neighbors among its k nearest neighbors in the frozen feature space, denoted by $\mathcal{N}_i^{\mathcal{A}}$, and aggregate their predictions as

$$q_i = \frac{1}{|\mathcal{N}_i^{\mathcal{A}}|} \sum_{j \in \mathcal{N}_i^{\mathcal{A}}} p_j. \tag{13}$$

Motivated by Theorem 2, CARE adapts the utilization strength according to reliability:

$$\alpha_i = \begin{cases} 0, & r_i \geq 1 - \tau, \\ 1 - r_i, & r_i < 1 - \tau, \end{cases} \qquad \widetilde{p}_i = (1 - \alpha_i) p_i + \alpha_i q_i. \tag{14}$$

The final prediction is $\widetilde{y}_i = \arg\max_k \widetilde{p}_{ik}$. Thus, reliable predictions are preserved, while less reliable samples receive stronger guidance from nearby reliable neighbors, allowing CARE to recover category-related information in the representation that is underutilized by the initial assignment. The contribution of each CARE component is further analyzed in Appendix F.

5 Experiment

5.1 Experimental Settings

Datasets and evaluation metrics. We conduct experiments on 6 widely used image clustering benchmarks, including ImageNet-10 and ImageNet-Dogs (Chang et al., 2017), STL-10 (Coates et al., 2011), CIFAR-10, CIFAR-20(the 20-superclass protocol of CIFAR-100), and CIFAR-100 (Krizhevsky et al., 2009), as well as 3 more challenging datasets, DTD (Cimpoi et al., 2014), UCF101 (Soomro et al., 2012), and ImageNet-1K (Deng et al., 2009). Detailed dataset descriptions are provided in Appendix A.1. We evaluate clustering performance using 3 widely adopted metrics: Accuracy (ACC) (Li & Ding, 2006), Normalized Mutual Information (NMI) (Estévez et al., 2009), and Adjusted Rand Index (ARI) (Steinley, 2004), where higher values indicate better performance.

Baselines. We apply CARE to two representative external-knowledge-assisted image clustering methods, TAC (Li et al., 2023) and SATC (Jiu et al., 2026), and compare the resulting methods

Table 1: Comparison of clustering performance (%) on 5 standard benchmarks. Within each paired backbone block, the better result is marked with an underline, while the overall best result is highlighted in **bold**.

Method	ImageNet-10			ImageNet-Dogs			STL-10			CIFAR-10			CIFAR-20		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
BYOL [NeurIPS'20]	94.7	88.4	88.9	72.9	69.7	60.9	86.1	75.4	71.5	87.5	78.0	75.2	52.3	53.3	36.0
PICA [CVPR'20]	87.0	80.2	76.1	35.2	35.2	20.1	71.3	61.1	53.1	69.6	59.1	51.2	33.7	31.0	17.1
GCC [ICCV'21]	90.1	84.2	82.2	52.6	49.0	36.2	78.8	68.4	63.1	85.6	76.4	72.8	47.2	47.2	30.5
IDFD [ICLR'21]	95.4	89.8	90.1	59.1	54.6	41.3	75.6	64.3	57.5	81.5	71.1	66.3	42.5	42.6	26.4
TCC [NeurIPS'21]	89.7	84.0	82.5	59.5	55.4	41.7	81.4	73.2	68.9	90.6	79.0	73.3	49.1	47.9	31.2
SPICE [TIP'22]	96.9	92.7	93.3	67.5	62.7	52.6	92.9	86.0	86.5	91.8	85.0	83.6	58.4	58.3	42.2
TCL [IJCV'22]	89.5	87.5	83.7	64.4	62.3	51.6	86.8	79.9	75.7	88.7	81.9	78.0	53.1	52.9	35.7
SACC [PR'23]	90.5	87.7	84.3	43.7	45.5	28.5	75.9	69.1	62.6	85.1	76.5	72.4	44.3	44.8	28.2
CoNR [NeurIPS'23]	96.4	91.1	92.2	79.4	74.4	66.7	92.6	85.2	84.6	93.2	86.7	86.1	60.4	60.4	44.3
DPAC [ECCV'24]	97.0	92.5	93.5	72.6	66.7	59.8	93.4	86.3	86.1	93.4	87.0	86.6	55.5	54.2	39.3
LFSS [ICML'25]	93.2	85.6	85.7	69.1	61.7	53.3	86.1	77.1	74.0	93.4	87.2	86.6	58.7	59.9	43.5
GradNorm [ICCV'25]	99.4	98.7	98.7	81.2	81.0	70.9	98.3	95.6	96.2	91.1	82.6	81.5	60.6	61.3	43.6
CAE [AAAI'26]	98.8	98.1	97.4	77.5	77.9	66.3	98.2	95.5	96.7	90.9	81.7	80.5	60.9	62.8	45.9
NTK-SC [ICLR'26]	99.2	97.8	98.4	84.9	82.4	71.4	98.3	95.8	96.3	92.0	83.3	83.0	59.6	63.3	43.5
TAC [ICML'24]	99.3	98.3	98.5	82.3	74.3	68.6	97.8	94.7	95.2	91.7	82.8	82.8	56.8	58.3	42.0
TAC + Ours	99.4	98.3	98.6	85.8	78.9	74.7	98.1	95.2	95.8	92.6	84.2	84.4	61.0	62.9	46.2
SATC [ICLR'26]	99.6	99.0	99.3	90.6	88.4	85.4	99.0	97.3	97.6	94.5	87.9	88.4	62.5	63.9	44.6
SATC + Ours	**99.7**	**99.1**	**99.4**	**91.1**	**89.3**	**86.2**	**99.1**	**97.4**	**97.8**	**94.7**	**88.1**	**88.6**	**64.5**	**65.5**	**47.5**

with a broad range of representative image clustering baselines, including BYOL (Grill et al., 2020), PICA (Huang et al., 2020), GCC (Zhong et al., 2021), IDFD (Tao et al., 2021), TCC (Shen et al., 2021), SPICE (Niu et al., 2022), TCL (Li et al., 2022), SACC (Deng et al., 2023), CoNR (Yu et al., 2023), DPAC (Yan et al., 2024), LFSS (Li et al., 2025b), GradNorm (Peng et al., 2025), CAE (Zhu et al., 2026), and NTK-SC (Peng et al., 2026).

Implementation Details. Following prior work (Li et al., 2023; Jiu et al., 2026), we use the frozen CLIP ViT-B/32 (Radford et al., 2021) as the visual backbone. For external knowledge learning, modality-specific clustering heads are implemented as 512-512-K MLPs and trained for 20 epochs with a batch size of 512 and $T = 0.5$ (for ImageNet-1K, 100 epochs, batch size 8192, and $T = 5.0$), $\lambda_1 = 1.0$, and $\lambda_2 = 5.0$. For assignment establishment, we freeze the knowledge-enhanced visual features and train the default linear clustering readout for 2,000 iterations, $\gamma = 10.0$, and a batch size of 10,000. For knowledge utilization, we use cosine kNN with $k = 20$ and $\tau = 0.5$, with parameter sensitivity analyses provided in Appendix I. Experiments are conducted on an NVIDIA H200 GPU, with a detailed runtime analysis provided in Appendix J. More implementation details are provided in Appendix A.2.

5.2 Comparison Experiment

Evaluation on standard datasets. As shown in Table 1, CARE consistently improves both TAC and SATC on 5 widely used datasets, demonstrating that better utilization of learned external knowledge benefits different knowledge-enhanced representations. For TAC, the gains are particularly notable on the fine-grained ImageNet-Dogs and CIFAR-20, where CARE improves ACC by 3.5% and 4.2%, respectively. Even on the nearly saturated ImageNet-10, CARE still yields stable improvements. We further observe a smaller gain on ImageNet-Dogs for SATC than for TAC, which may be because SATC's spatial knowledge is already more effectively utilized, leaving less room for CARE to recover. Additional results on the full CIFAR-100 benchmark with all 100 fine-grained classes are provided in Appendix B.

Evaluation on challenging datasets. As shown in Table 2, CARE yields more pronounced improvements on the three challenging datasets. For TAC, CARE improves the average ACC/NMI/ARI from 57.9/73.3/44.8 to 61.2/75.4/48.6, with ACC gains of 4.1% on DTD and 4.6% on UCF-101. CARE also consistently improves SATC, increasing the average ACC/NMI/ARI from 62.6/76.0/48.5

Table 2: Comparison of clustering performance (%) on 3 challenging benchmarks. Within each paired backbone block, the better result is marked with an underline, while the overall best result is highlighted in **bold**.

Method	DTD			UCF-101			ImageNet-1K			Average		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
GradNorm [ICCV'25]	50.9	63.1	34.2	62.7	82.9	53.2	52.6	79.2	39.1	55.4	75.1	42.2
CAE [AAAI'26]	46.7	62.1	31.9	63.5	82.8	54.9	53.1	79.3	39.3	54.4	74.7	42.0
NTK-SC [ICLR'26]	52.0	61.7	33.6	67.9	83.0	59.4	56.3	79.2	39.4	58.7	74.6	44.1
TAC [ICML'24]	48.4	58.1	31.7	68.3	81.9	59.7	57.1	79.9	43.0	57.9	73.3	44.8
TAC + Ours	52.5	61.3	36.5	72.9	82.8	64.2	58.2	82.1	45.0	61.2	75.4	48.6
SATC [ICLR'26]	53.0	62.5	37.0	70.7	82.6	62.0	64.0	83.0	46.5	62.6	76.0	48.5
SATC + Ours	**54.9**	**63.7**	**39.0**	**74.2**	**84.0**	**65.9**	**65.2**	**84.0**	**46.7**	**64.8**	**77.2**	**50.5**

to 64.8/77.2/50.5. Compared with the standard benchmarks, these larger gains demonstrate the greater benefit of improving knowledge utilization in more complex clustering scenarios.

5.3 Mechanism Validation of Knowledge Utilization

Existence of the Learning-Utilization Gap. To examine whether the *learning-utilization gap* is a general phenomenon and how it changes after incorporating external knowledge, we evaluate raw CLIP, TAC, and SATC representations on 9 datasets. For each frozen representation, we define the gap as the ACC difference between the cluster assignments produced by the clustering readout and an analysis-only kNN oracle. As shown in Fig. 3, external knowledge learning reduces the average gap from 19.06% for raw CLIP to 9.37% for TAC and 9.12% for SATC. However, substantial residual gaps remain, especially on challenging datasets such as ImageNet-1K, indicating that the learned representations still contain category-related information that is not fully reflected in the final cluster assignments.

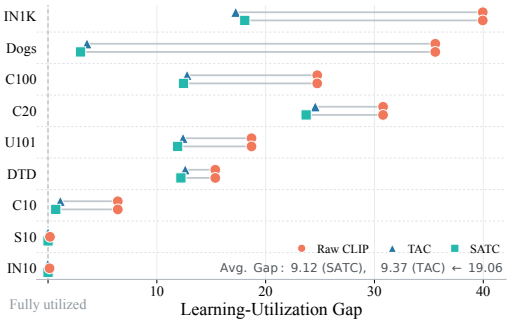
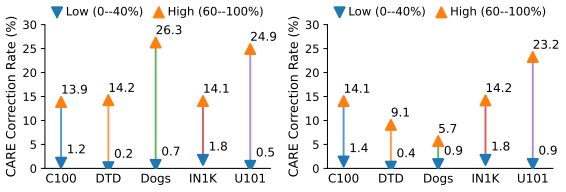


Figure 3: Learning-Utilization gap.

Structural Support Facilitates Error Recovery. To examine whether CARE recovers errors using category-related structure already encoded in the frozen representation, we analyze initially misclustered samples on 5 representative datasets. For each sample, structural support is measured as the proportion of its k nearest neighbors sharing the same ground-truth category, with labels used only for post-hoc analysis. We compare low-support (0–40%) and high-support (60–100%) samples, and report the proportion of initially misclustered samples that are successfully corrected by CARE. As shown in Fig. 4, high-support samples achieve substantially higher correction rates for both TAC and SATC, supporting our theoretical finding that reliable local structure can recover underutilized knowledge.



(a) TAC (b) SATC

Figure 4: Structural support facilitates error recovery.

5.4 Generality across External Knowledge

To examine whether CARE can effectively utilize different forms of external knowledge, we evaluate it with three representative knowledge sources: predefined-vocabulary textual features, LLM-generated textual features, and spatial structural features. Experiments are conducted on CIFAR-10, CIFAR-100, and ImageNet-Dogs. For each knowledge source, we keep the corresponding knowledge-enhanced representation fixed and apply CARE using the same configuration.

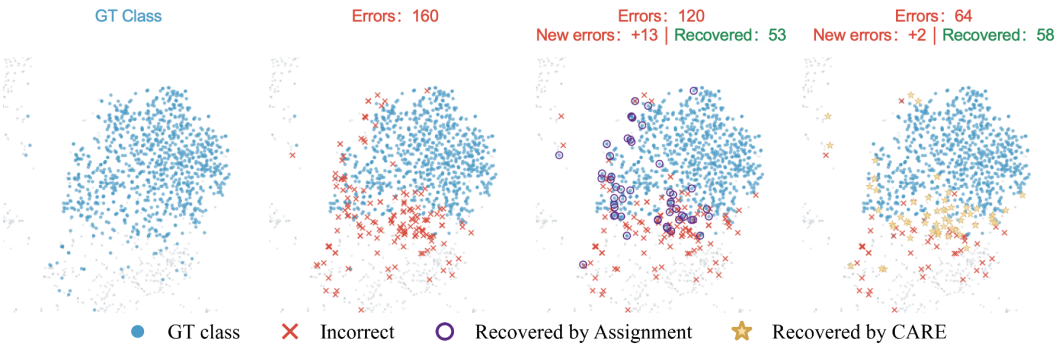


Figure 5: Visualization of assignment recovery on ImageNet-Dogs.

As shown in Table 3, CARE consistently improves clustering performance across all three forms of external knowledge. These results demonstrate that CARE is not tied to a specific form of external knowledge, but provides a general mechanism for better utilizing knowledge already encoded in learned representations.

5.5 Visualization

To qualitatively examine how CARE progressively utilizes knowledge already encoded in the learned representation, we visualize one representative cluster from ImageNet-Dogs dataset using t-SNE. Blue points denote samples from the target ground-truth class, red crosses denote incorrect assignments, purple circles indicate errors recovered by Assignment Establishment, and orange stars indicate additional errors recovered by the complete CARE framework.

Table 3: Generality of CARE across different external knowledge sources. Within each paired backbone block, the better result is marked with an underline.

Knowledge Source	CIFAR-10			CIFAR-100			ImageNet-Dogs		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Lexicon Text	91.7	82.8	82.8	57.5	67.9	41.7	82.3	74.3	68.6
Lexicon Text + Ours	92.6	84.2	84.4	61.2	68.3	47.0	85.8	78.9	74.7
LLM Text	90.5	81.1	80.4	52.5	62.9	38.1	49.7	46.0	33.6
LLM Text + Ours	91.7	83.1	82.8	60.4	67.9	46.2	54.0	51.0	38.7
Spatial Priors	93.6	87.2	86.9	55.8	67.7	42.4	90.6	88.4	85.4
Spatial Priors + Ours	93.7	87.9	87.3	57.8	68.6	44.3	91.1	89.3	86.2

As shown in Fig. 5, Assignment Establishment translates part of the learned external knowledge into more reliable cluster assignments, reducing the number of errors from 160 to 120 by correcting 53 samples, although it also introduces 13 new errors. The complete CARE further reduces the errors to 64 by correcting 58 additional samples while introducing only 2 new errors. Most recovered samples lie in locally coherent class structures, showing that CARE more effectively translates learned external knowledge into reliable cluster assignments with fewer newly introduced errors.

6 Conclusion

In this work, we address an underexplored problem in external-knowledge-assisted image clustering: existing methods mainly focus on acquiring, selecting, or transferring useful external knowledge into image representations, while largely assuming that the learned knowledge will be effectively reflected in the final cluster assignments. We identify this mismatch as the learning-utilization gap and provide an information-theoretic characterization showing that useful category-related knowledge may already be encoded in learned representations but remain insufficiently utilized by final predictions. We further theoretically establish when distance-based structural relations can recover such underutilized knowledge and correct erroneous predictions. Guided by these theoretical insights, we propose CARE, which establishes stable cluster assignments, assesses sample-level reliability, and adaptively exploits local feature-space structure to refine unreliable predictions. Extensive experiments across standard and challenging benchmarks, different external knowledge sources, and different clustering readouts consistently demonstrate the effectiveness and generality of CARE. Future work will explore more adaptive knowledge-utilization strategies and extend CARE to broader clustering and multimodal learning scenarios.

AI USE STATEMENT

In this work, we used generative AI tools only for manuscript writing assistance, including improving clarity and readability. We did not use generative AI tools to develop the theoretical framework, construct the proposed algorithm, design experiments, generate experimental results, or make scientific conclusions. All AI-assisted text was carefully reviewed and revised by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed CARE framework, optimization objectives, and implementation settings in the main paper. Complete proofs of the theoretical results are provided in the appendix, together with additional experimental details, dataset configurations, and feature-generation procedures. We also report all major hyperparameters and evaluation protocols used in the experiments. The source code will be released upon publication to facilitate reproduction of the reported results.

REFERENCES

Robert A Bjork, Elizabeth L Bjork, et al. A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes*, 2:35–67, 1992.

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, pp. 6869–6878, 2023.

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*, pp. 5879–5887, 2017.

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*, pp. 3606–3613, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Thomas M Cover, Joy A Thomas, and John Kieffer. *Elements of information theory*, volume 2. wiley New York, 1991.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Xiaozhi Deng, Dong Huang, Ding-Hua Chen, Chang-Dong Wang, and Jian-Huang Lai. Strongly augmented contrastive clustering. *Pattern Recognition*, 139:109470, 2023.

Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. Normalized mutual information feature selection. *IEEE Transactions on neural networks*, 20(2):189–201, 2009.

Artyom Gadetsky and Maria Brbic. The pursuit of human labeling: a new perspective on unsupervised learning. *Advances in Neural Information Processing Systems*, 36:60527–60546, 2023.

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. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. pp. 8846 – 8855. IEEE, 2020. ISBN 9781728171685. doi: 10.1109/CVPR42600.2020.00887.

Zizheng Jiu, Feijiang Li, Jieting Wang, Yuhua Qian, and Lu Chen. Spatial structure and selective text jointly facilitate image clustering. In *International Conference on Learning Representations*, volume 2026, pp. 155690–155717, 2026.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Feijiang Li, Zhenxiong Li, Jieting Wang, Zizheng Jiu, Saixiong Liu, and Liang Du. Reducing bias and variance: Generative semantic guidance and bi-layer ensemble for image clustering. *arXiv preprint arXiv:2605.12961*, 2026.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 14066–14091. Curran Associates, Inc., 2025a. doi: 10.52202/085713-0473.

Tao Li and Chris Ding. The relationships among various nonnegative matrix factorization methods for clustering. In *Sixth International Conference on Data Mining (ICDM'06)*, pp. 362–371. IEEE, 2006.

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.

Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. *arXiv preprint arXiv:2310.11989*, 2023.

Zhixin Li, Yuheng Jia, Hui Liu, and Junhui Hou. Learning from sample stability for deep clustering. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34904–34919. PMLR, 13–19 Jul 2025b.

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

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

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*, pp. 19805–19815, 2025.

Bo Peng, Yuanwei Hu, Bo Liu, Ling Chen, Jie Lu, and Zhen Fang. Delving into spectral clustering with vision-language representations. In *The Fourteenth International Conference on Learning Representations*, 2026.

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*, pp. 8748–8763. PmLR, 2021.

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

Nicholas C Soderstrom and Robert A Bjork. Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2):176–199, 2015.

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.

Douglas Steinley. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3): 386–396, 2004.

Yaling Tao, Kentaro Takagi, and Kouta Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. *arXiv preprint arXiv:2106.00131*, 2021.

Endel Tulving and Zena Pearlstone. Availability versus accessibility of information in memory for words. *Journal of verbal learning and verbal behavior*, 5(4):381–391, 1966.

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*, pp. 268–285. Springer, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Yuxuan Yan, Na Lu, and Ruofan Yan. Deep online probability aggregation clustering. In *European Conference on Computer Vision*, pp. 37–54. Springer, 2024.

Chunlin Yu, Ye Shi, and Jingya Wang. Contextually affinitive neighborhood refinery for deep clustering. *Advances in Neural Information Processing Systems*, 36:5778–5790, 2023.

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 (ICCV)*, pp. 9224–9233, October 2021.

Yijie Zhong, Yunfan Gao, Weipeng Jiang, and Haofen Wang. Hierarchical textual knowledge for enhanced image clustering. *arXiv preprint arXiv:2604.11144*, 2026.

Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16, pp. 321–328. MIT Press, 2003.

Xingyu Zhu, Beier Zhu, Yunfan Li, Junfeng Fang, Shuo Wang, Kesen Zhao, and Hanwang Zhang. Hierarchical semantic alignment for image clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29177–29185, 2026.

Imtiaz Ziko, Jose Dolz, Eric Granger, and Ismail Ben Ayed. Laplacian regularized few-shot learning. In *International conference on machine learning*, pp. 11660–11670. PMLR, 2020.

APPENDIX

The appendix is organized as follows:

- A Experimental Settings
- B Evaluation on CIFAR-100
- C Proofs of Theoretical Results
- D Stable Assignment Establishment
- E Reliability of Selected Anchors
- F Component Ablation
- G Generality across Clustering Readouts
- H Generality to Multiple Auxiliary Modalities
- I Parameter Sensitivity
- J Time Analysis

A EXPERIMENTAL SETTINGS

A.1 DATASETS

We evaluate CARE on 9 image clustering benchmarks. The 6 widely used benchmarks include CIFAR-10, CIFAR-20, and CIFAR-100 (Krizhevsky et al., 2009), STL-10 (Coates et al., 2011), ImageNet-10, and ImageNet-Dogs (Chang et al., 2017). To further evaluate the scalability and generalization ability of CARE, we also consider 3 more challenging datasets: DTD (Cimpoi et al., 2014), UCF101 (Soomro et al., 2012), and ImageNet-1K (Deng et al., 2009). DTD is a texture recognition dataset, UCF101 contains human action videos, and ImageNet-1K provides a large-scale clustering setting with 1,000 semantic categories. For UCF101, we use the middle frame of each video clip as the input to the pretrained image encoders. Detailed statistics of all datasets are summarized in Table 4.

Table 4: Summary statistics of the 9 datasets used in our experiments.

Dataset	Num. Classes	Train Size	Test Size
CIFAR-10 (Krizhevsky et al., 2009)	10	50,000	10,000
CIFAR-20 (Krizhevsky et al., 2009)	20	50,000	10,000
CIFAR-100 (Krizhevsky et al., 2009)	100	50,000	10,000
STL-10 (Coates et al., 2011)	10	5,000	8,000
ImageNet-10 (Chang et al., 2017)	10	10,500	2,630
ImageNet-Dogs (Chang et al., 2017)	15	15,600	3,900
DTD (Cimpoi et al., 2014)	47	3,760	1,880
UCF101 (Soomro et al., 2012)	101	9,537	3,783
ImageNet-1K (Deng et al., 2009)	1,000	1,281,167	50,000

A.2 IMPLEMENTATION DETAILS

kNN Oracle. We use an analysis-only kNN oracle to evaluate the class-consistent structure preserved in frozen representations. We ℓ_2-normalize the features and retrieve $k = 20$ nearest neighbors using cosine similarity, excluding the query itself, and predict each sample by majority voting over neighbor ground-truth labels. The oracle is independently evaluated on Raw CLIP, TAC, and SATC features. Ground-truth labels are used only for post-hoc analysis and are never accessed by CARE.

Predefined-vocabulary textual features. Following TAC (Li et al., 2023), we collect candidate nouns from WordNet (Miller, 1995) and encode them with the CLIP ViT-B/32 text encoder. The visual features are first partitioned into $\lfloor N/300 \rfloor$ clusters using K-Means. For each cluster, the top-5

confident nouns are selected according to their similarity to the corresponding visual center. The textual feature of each image is then constructed through soft retrieval over the selected nouns with temperature 0.005.

LLM-generated textual features. Following GSEC (Li et al., 2026), we first extract visual features using CLIP ViT-B/32 and perform K-Means to uncover coarse semantic groups, where the number of clusters is set to $\max\{n/300, 3K\}$, with n and K denoting the number of samples and target clusters, respectively. Within each cluster, images are sorted by their distance to the cluster center, and up to 5 representative images are uniformly sampled along the sorted order. We then employ Llama-3.2-Vision-11B to generate a semantic description for each representative image using the prompt: *"Identify and describe the main object in this image. Respond with the format: This image contains a [object] characterized by [attribute1], [attribute2], and [attribute3]."* The generated descriptions are encoded by the CLIP ViT-B/32 text encoder. For each image, we compute its cosine similarities to all generated text embeddings and normalize them using a temperature-scaled softmax with a temperature of 0.005. The final image-specific textual feature is obtained as the weighted average of these text embeddings. In our implementation, generation is limited to 80 output tokens.

Spatial structural features. Following SATC (Jiu et al., 2026), each image is divided into patches and patch-level features are extracted using a pretrained ResNet-50 with the final pooling and classification layers removed. The patch features are treated as graph nodes and processed by a Graph Attention Network (GAT) (Veličković et al., 2017) to capture relations among different image regions. Global average pooling over the refined patch representations produces the spatial feature. For the complete SATC pipeline, we follow its original textual selection strategy and set the compactness threshold to 0.33.

B Evaluation on CIFAR-100

To further evaluate CARE under a more challenging clustering setting, we conduct additional experiments on the full CIFAR-100 benchmark using all 100 fine-grained classes. Unlike the commonly used CIFAR-20 protocol, which evaluates clustering on the 20 superclasses of CIFAR-100, this setting directly uses all 100 original categories.

Table 5: Clustering performance (%) on CIFAR-100 using all 100 fine-grained classes. Within each paired backbone block, the better result is marked with an underline, while the overall best result is highlighted in **bold**.

Method	ACC	NMI	ARI
TAC	57.5	67.9	41.7
TAC + Ours	<u>61.2</u>	<u>68.3</u>	<u>47.0</u>
SATC	60.7	69.9	46.4
SATC + Ours	**63.8**	**70.9**	**49.8**

As shown in Table 5, CARE consistently improves both TAC and SATC under the full 100-class setting. Specifically, CARE improves TAC by 3.7%, 0.4%, and 5.3% in ACC, NMI, and ARI, respectively, while improving SATC by 3.1%, 1.0%, and 3.4%. These results further demonstrate that CARE remains effective when the number of clusters increases substantially, indicating its ability to better utilize learned external knowledge under more challenging fine-grained clustering settings.

C Proofs of Theoretical Results

C.1 Proof of Theorem 1

Proof 1 *Since $P = h(Z)$ is deterministic, $H(Y \mid Z, P) = H(Y \mid Z)$. Therefore,*

$$I(Y; Z \mid P) = H(Y \mid P) - H(Y \mid Z, P) \tag{15}$$
$$= H(Y \mid P) - H(Y \mid Z), \tag{16}$$

which yields

$$H(Y \mid P) = H(Y \mid Z) + I(Y; Z \mid P). \tag{17}$$

By the non-negativity of conditional mutual information, $H(Y \mid P) \geq H(Y \mid Z)$, with equality if and only if $I(Y; Z \mid P) = 0$. Thus, a positive gap indicates that Z contains category-related information that is not fully recoverable from P. This completes the proof.

C.2 Proof of Theorem 2

Proof 2 *For any competing category $c \neq y_i$, Eq. (4) gives*

$$\widetilde{p}_{iy_i} - \widetilde{p}_{ic} = (1-\alpha_i)(p_{iy_i} - p_{ic}) + \alpha_i(q_{iy_i} - q_{ic}) \tag{18}$$

$$= -(1-\alpha_i)\delta_{ic} + \alpha_i\gamma_{ic}. \tag{19}$$

If $\delta_{ic} > 0$, then $\gamma_{ic} > 0$ implies

$$\widetilde{p}_{iy_i} - \widetilde{p}_{ic} > 0 \quad \Longleftrightarrow \quad \alpha_i > \frac{\delta_{ic}}{\delta_{ic} + \gamma_{ic}}. \tag{20}$$

If $\delta_{ic} \leq 0$, the margin in Eq. (19)is positive for every $\alpha_i \in (0, 1]$.

Consequently,

$$\alpha_i > \max_{c \neq y_i} \frac{[\delta_{ic}]_+}{[\delta_{ic}]_+ + \gamma_{ic}}, \qquad [x]_+ = \max(x, 0), \tag{21}$$

ensures $\widetilde{p}_{iy_i} > \widetilde{p}_{ic}$ for all $c \neq y_i$. Hence, $\widetilde{p}_i$ uniquely predicts y_i.

Furthermore, for $\delta_{ic} > 0$, let $\tau_{ic} = \delta_{ic}/(\delta_{ic} + \gamma_{ic})$. Then

$$\frac{\partial \tau_{ic}}{\partial \delta_{ic}} = \frac{\gamma_{ic}}{(\delta_{ic} + \gamma_{ic})^2} > 0, \qquad \frac{\partial \tau_{ic}}{\partial \gamma_{ic}} = -\frac{\delta_{ic}}{(\delta_{ic} + \gamma_{ic})^2} < 0. \tag{22}$$

Thus, a larger prediction deficit increases the required adjustment, whereas a larger structural margin decreases it. Since every threshold in Eq. (5) lies in $[0, 1)$, a feasible correction coefficient exists. This completes the proof.

C.3 Proof of Corollary 1

Proof 3 *For each $c \neq y_i$, the assumed neighbor margin bounds are*

$$p_{jy_i} - p_{jc} \geq \begin{cases} \mu_c^+, & y_j = y_i, \\ -\mu_c^-, & y_j \neq y_i, \end{cases} \qquad \mu_c^+ > 0, \quad \mu_c^- \geq 0. \tag{23}$$

Since a fraction ρ_i of the selected neighbors share category y_i,

$$\gamma_{ic} = \frac{1}{|\mathcal{N}_i^+|} \sum_{j \in \mathcal{N}_i^+} (p_{jy_i} - p_{jc}) \tag{24}$$

$$\geq \rho_i \mu_c^+ - (1-\rho_i)\mu_c^- \tag{25}$$

$$= (\mu_c^+ + \mu_c^-)\rho_i - \mu_c^-. \tag{26}$$

Substituting into Eq. (19) yields

$$\widetilde{p}_{iy_i} - \widetilde{p}_{ic} \geq L_{ic}(\rho_i) := -(1-\alpha_i)\delta_{ic} + \alpha_i\big[(\mu_c^+ + \mu_c^-)\rho_i - \mu_c^-\big]. \tag{27}$$

Holding p_i, α_i, and the margin bounds fixed,

$$\frac{\partial L_{ic}}{\partial \rho_i} = \alpha_i(\mu_c^+ + \mu_c^-) \geq 0, \tag{28}$$

with strict positivity when $\alpha_i > 0$. Hence, increasing ρ_i raises the lower bound on each refined margin.

For $\alpha_i > 0$, requiring $L_{ic}(\rho_i) > 0$ for all $c \neq y_i$ gives the sufficient correction condition

$$\rho_i > \max_{c \neq y_i} \frac{\mu_c^- + \frac{1-\alpha_i}{\alpha_i}\delta_{ic}}{\mu_c^+ + \mu_c^-}. \tag{29}$$

Thus, stronger class-consistent support improves the sufficient correction guarantee under the fixed quantities above. This completes the proof.

Table 6: Effect of assignment establishment. Time denotes the runtime of the Assignment Establishment stage after obtaining the frozen TAC-enhanced visual representation. The overall best result is highlighted in **bold**.

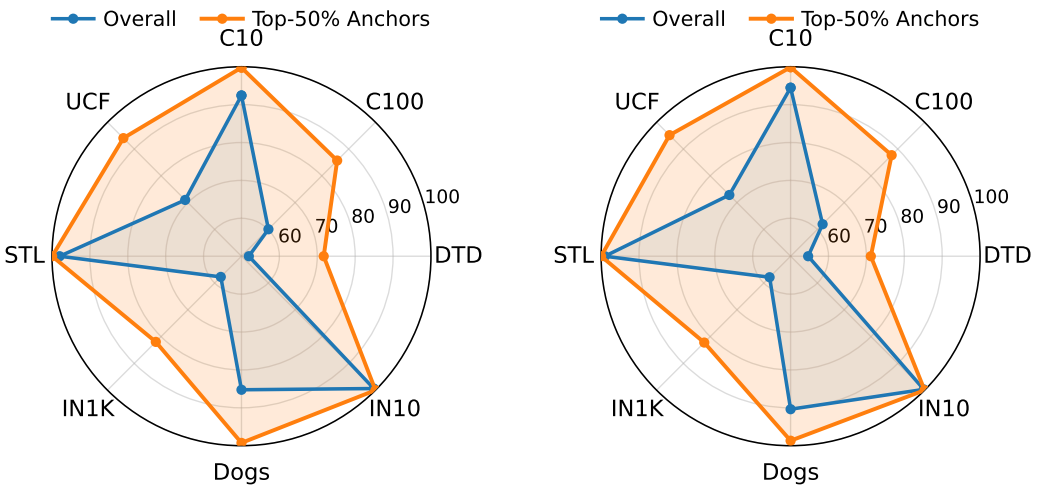
Iterations	CIFAR-100				ImageNet-Dogs				UCF-101			
	ACC	NMI	ARI	Time (s)	ACC	NMI	ARI	Time (s)	ACC	NMI	ARI	Time (s)
0	17.2	37.2	7.3	0.0	28.8	30.1	12.8	0.0	28.0	56.5	15.9	0.0
500	58.1	67.3	44.3	16.4	85.0	78.6	74.4	13.1	72.0	82.5	63.6	6.5
1000	60.1	67.9	45.6	32.3	85.2	78.7	74.5	26.5	72.6	82.6	64.1	11.9
2000	61.2	68.3	47.0	63.9	85.8	78.9	74.7	56.6	**72.9**	**82.8**	**64.2**	22.7
4000	**61.3**	**68.4**	**47.2**	126.8	**85.9**	**79.0**	**74.9**	87.9	72.7	82.7	63.9	44.2

D Stable Assignment Establishment

To analyze the effect of stable assignment establishment, we first perform TAC-based external knowledge learning with textual knowledge and obtain the corresponding knowledge-enhanced visual representations. We then freeze these visual features and perform Assignment Establishment with different numbers of optimization iterations on CIFAR-100, ImageNet-Dogs, and UCF-101. The 0-iteration setting corresponds to the randomly initialized cluster readout before any assignment optimization. For each setting, we subsequently apply the remaining CARE stages and report the final clustering performance of the complete CARE framework, while the reported runtime measures only the Assignment Establishment stage under the corresponding number of optimization iterations.

As shown in Table 6, clustering performance improves rapidly during the early optimization stage and gradually saturates as the number of iterations increases. This indicates that Assignment Establishment progressively translates part of the category-related information already encoded in the frozen knowledge-enhanced representation into more reliable cluster assignments. Increasing the number of iterations from 2,000 to 4,000 yields only marginal performance gains while substantially increasing the runtime. We therefore use 2,000 iterations by default as a practical trade-off between clustering performance and computational cost.

E Reliability of Selected Anchors



(a) TAC (b) SATC

Figure 6: Reliability of selected anchors.

To evaluate whether CARE's label-free reliability assessment can identify reliable predictions, we rank samples according to the soft neighborhood-support score used by CARE and select the top

50% as anchors. Ground-truth labels are used only for post-hoc evaluation. We compare the prediction accuracy of these selected anchors with the overall assignment accuracy across 8 datasets for both TAC and SATC.

As shown in Fig. 6, the selected anchors consistently achieve higher prediction accuracy than the overall sample set across all TAC and SATC settings. Across all 16 settings, the top-50% anchors achieve an average accuracy above 90%. These results show that CARE's soft neighborhood-support score can reliably concentrate correct predictions into the selected anchor set, providing an effective observable proxy for the correctly predicted neighbors assumed in Theorem 2.

F Component Ablation

Table 7: Full factorial component analysis of CARE on TAC. The overall best result is highlighted in **bold**.

Assignment Training	Anchor Selection	Adaptive Blending	CIFAR-100			ImageNet-Dogs			UCF-101		
			ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
			57.5	67.9	41.7	82.3	74.3	68.6	68.3	81.9	59.7
✓			60.2	67.6	45.6	84.5	77.2	72.6	70.9	82.5	61.2
	✓		59.0	67.3	43.4	83.8	76.6	71.5	69.4	80.8	59.3
		✓	59.2	67.6	43.6	84.2	77.1	72.2	69.6	81.2	59.7
	✓	✓	59.4	67.8	43.9	84.3	77.1	72.3	69.9	81.5	59.8
✓	✓		61.0	68.2	46.9	85.6	78.7	74.6	71.6	82.1	62.7
✓		✓	60.9	68.2	46.5	85.6	78.8	74.6	71.8	82.3	62.9
✓	✓	✓	**61.2**	**68.3**	**47.0**	**85.8**	**78.9**	**74.7**	**72.9**	**82.8**	**64.2**

To investigate the contribution of each component in CARE, we conduct a full factorial ablation study on TAC representation using CIFAR-100, ImageNet-Dogs, and UCF-101. We independently enable or disable three components: assignment training, reliable anchor selection, and adaptive blending. When anchor selection is enabled without adaptive blending, all non-anchor samples use a fixed blending coefficient of $\alpha = 0.75$, while adaptive blending follows the sample-specific reliability weighting defined in CARE.

As shown in Table 7, each component contributes to the final performance, while their combination consistently achieves the best results across all three datasets. Assignment training provides a strong improvement over the original predictions, indicating that establishing a more suitable cluster assignment is an important prerequisite for subsequent knowledge utilization. Reliable anchor selection and adaptive blending also improve performance when applied individually, showing that both restricting knowledge sources to reliable samples and adjusting the utilization strength are beneficial. More importantly, combining assignment training with either component yields further gains, and the complete CARE consistently performs best. These results confirm that CARE benefits from the complementary effects of assignment establishment, reliability-aware anchor selection, and adaptive knowledge utilization rather than relying on any single component.

G Generality across Clustering Readouts

Table 8: Generality of CARE across different clustering readouts. Within each paired backbone block, the better result is marked with an underline.

Readout	CIFAR-100			ImageNet-Dogs			UCF-101		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Linear	60.2	67.6	45.6	84.5	77.2	72.6	70.9	82.5	61.2
Linear + Ours	<u>61.2</u>	<u>68.3</u>	<u>47.0</u>	<u>85.8</u>	<u>78.9</u>	<u>74.7</u>	<u>72.9</u>	<u>82.8</u>	<u>64.2</u>
2-layer MLP	60.3	68.4	45.5	84.5	76.9	72.4	67.2	81.0	57.6
2-layer MLP + Ours	<u>61.5</u>	<u>68.5</u>	<u>47.2</u>	<u>85.9</u>	<u>79.0</u>	<u>74.8</u>	<u>71.0</u>	<u>82.1</u>	<u>62.3</u>
3-layer MLP	56.8	67.1	42.7	82.5	77.0	71.8	68.7	81.2	58.4
3-layer MLP + Ours	<u>58.6</u>	<u>67.4</u>	<u>45.2</u>	<u>83.8</u>	<u>78.8</u>	<u>74.3</u>	<u>72.2</u>	<u>82.5</u>	<u>63.4</u>

To examine whether the effectiveness of CARE depends on a specific clustering readout, we first perform TAC-based external knowledge learning with textual knowledge and obtain the corresponding knowledge-enhanced visual representations. These representations are then frozen, and we replace the default linear clustering readout in Assignment Establishment with two nonlinear alternatives, namely a 2-layer MLP and a 3-layer MLP. Only the outer clustering readout is changed, while the inner learner remains a freshly initialized linear model at every iteration. All other settings remain unchanged.

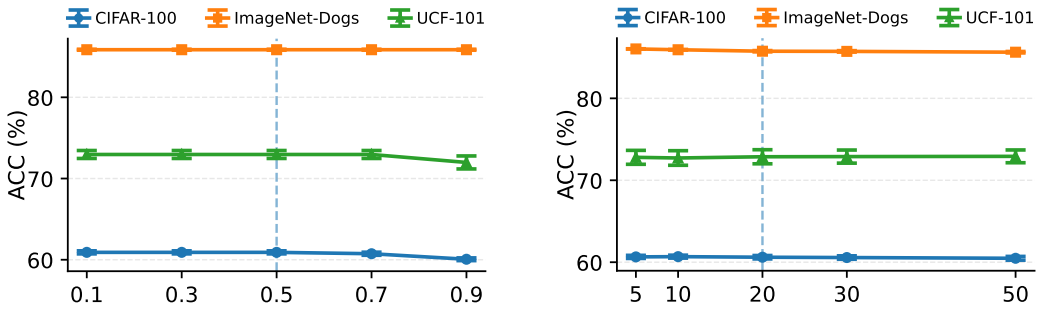
As shown in Table 8, CARE consistently improves the corresponding baseline readout across all three datasets. Notably, the linear readout already provides strong initial cluster assignments, indicating that external knowledge learning has produced a feature space with substantial category-level separability. This observation suggests that part of the learned external knowledge can already be translated into reliable cluster assignments through a simple linear readout, while the subsequent reliability assessment and knowledge utilization stages further recover category-related structure that remains underutilized. Moreover, the consistent improvements obtained with both 2-layer and 3-layer MLP readouts show that CARE is not restricted to a particular readout architecture.

H Generality to Multiple Auxiliary Modalities

To evaluate whether CARE remains effective when external knowledge learning involves multiple auxiliary modalities, we further analyze the three-modal settings of SATC. In SATC, textual knowledge is selectively retained on CIFAR-10, STL-10, CIFAR-100, and ImageNet-1K, where visual, textual, and spatial modalities are jointly used during external knowledge learning.

The corresponding SATC and SATC+Ours results are reported in Tables 1, Tables 2, and Table 5. CARE consistently improves SATC under these three-modal settings. This demonstrates that CARE is not restricted to external knowledge learning with a single auxiliary modality and can also effectively utilize visual representations enhanced by multiple complementary auxiliary modalities.

I Parameter Sensitivity



(a) Anchor ratio τ (b) Neighborhood size k

Figure 7: Parameter sensitivity of CARE to the anchor ratio τ and neighborhood size k.

We analyze the sensitivity of CARE to its two main hyperparameters, the reliable-anchor ratio τ and neighborhood size k, on CIFAR-100, ImageNet-Dogs, and UCF-101. We first perform TAC-based external knowledge learning with textual knowledge and obtain the corresponding knowledge-enhanced visual representations, which are then frozen throughout the sensitivity analysis. We vary $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $k \in \{5, 10, 20, 30, 50\}$ while keeping all other settings fixed. All experiments are repeated over 5 random seeds.

I.1 Anchor ratio τ.

As shown in Fig. 7(a), CARE remains highly stable when τ varies from 0.1 to 0.5. After external knowledge learning and Assignment Establishment, a large fraction of samples already exhibits

structurally consistent and correct cluster assignments. Therefore, selecting different subsets from the most reliable samples produces similar anchor sets and leads to stable performance within this range. However, an overly large τ includes less reliable anchors while leaving fewer samples to be corrected. This reduces both the reliability and coverage of structural guidance, leading to performance degradation.

I.2 NEIGHBORHOOD SIZE k

Fig. 7(b) shows that CARE is insensitive to a wide range of neighborhood sizes. This robustness arises because structural evidence is aggregated only from reliable anchors in the locally coherent knowledge-enhanced feature space. Consequently, moderate changes in k largely preserve the dominant semantic structure while averaging out individual noisy predictions. A very small neighborhood may provide insufficient structural evidence, whereas an excessively large neighborhood may introduce semantically less relevant samples and increase the computational cost of neighborhood retrieval and aggregation. As further reflected by the stage-wise runtime in Table 9, neighborhood-based Reliability Assessment and Knowledge Utilization introduce noticeable overhead on large-scale datasets, making an unnecessarily large k undesirable. The stable performance across $k \in [5, 50]$ therefore indicates that CARE does not require a carefully tuned or excessively large neighborhood. We use $k = 20$ by default as a practical trade-off between sufficient structural evidence and computational efficiency.

Based on these results, we use $\tau = 0.5$ and $k = 20$ throughout all experiments, providing a practical balance between anchor reliability and sufficient local structural evidence.

J TIME ANALYSIS

Table 9: Runtime breakdown of the complete external-knowledge learning and utilization pipeline. All times are reported in seconds on a single NVIDIA H200 GPU.

Dataset	Knowledge Learning	Assignment Establishment	Reliability Assessment	Knowledge Utilization	Total Time
CIFAR-100	100.98	63.86	0.71	0.03	165.58
ImageNet-Dogs	36.42	56.59	0.19	0.01	93.21
UCF-101	20.22	22.66	0.17	0.01	43.06
ImageNet-1K	2629.42	125.26	216.07	10.25	2981.00

To analyze the computational cost of the complete pipeline, we separately report the runtime of its four stages: External Knowledge Learning, Assignment Establishment, Reliability Assessment, and Knowledge Utilization. Specifically, we first perform TAC-based external knowledge learning with textual knowledge to obtain knowledge-enhanced visual representations, which are then frozen and used throughout the subsequent CARE stages.

As shown in Table 9, the computational cost varies across datasets and stages. External Knowledge Learning dominates the overall runtime on large-scale ImageNet-1K, while Assignment Establishment represents the main subsequent cost on CIFAR-100, ImageNet-Dogs, and UCF-101. Reliability Assessment and Knowledge Utilization are generally inexpensive, although the former becomes more costly on ImageNet-1K due to large-scale neighborhood construction. Overall, CARE introduces additional but manageable computational cost after obtaining the knowledge-enhanced representation.