STRUCTURE SHAPES SEMANTICS: A UNIFIED FRAMEWORK FOR TEXT-ASSISTED IMAGE CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-assisted Image Clustering (TIC) leverages vision-language models to provide external textual semantic information, thereby enhancing clustering performance. Existing approaches typically follow a two-stage paradigm: semantic representation construction followed by clustering network training. However, these methods heavily rely on handcrafted strategies or point-to-point matching for semantic representation construction, making them vulnerable to imperfect image-text alignment and semantic noise. In this work, we first introduce a principled framework that reformulates semantic representation construction as a global optimization problem to address these challenges, and propose to solve it via coordinate descent. Then, guided by the principle that "Structure Shapes Semantics" (S^3), we introduce several semantic-quality criteria and formulate them as optimization objectives. Building upon this reformulation, we further unify the two TIC stages under the shared objective and reveal their symmetric optimization processes. To further refine the semantic representations, we propose the Fine-Structure guided Optimization (FSO). Extensive experiments validate the effectiveness of our method, with a 6.0% accuracy gain on the most challenging Tiny-ImageNet. We believe this work advances TIC from empirically driven heuristics toward a principled optimization paradigm. Code will be released.

1 INTRODUCTION

Image clustering (IC) (MacQueen et al., 1967; Ester et al., 1996; Gowda & Krishna, 1978; Zelnik-Manor & Perona, 2005; Nie et al., 2011) seeks to group unlabeled images into semantically coherent clusters. Deep IC has progressed from feature learning followed by clustering to end-to-end generative, contrastive, and neighborhood-consistency frameworks (Coates & Ng, 2012; Prasad et al., 2020; Mrabah et al., 2020; Li et al., 2019; 2021b; Zhong et al., 2021; Li et al., 2022c). However, the reliance on intrinsic image information inherently limits performance. Recently, advances in multi-modal pretraining (Radford et al., 2021; Lin et al., 2023; Li et al., 2022b) have catalyzed remarkable progress across numerous visual tasks (Liu et al., 2023; Kirillov et al., 2023). Building on this momentum, a suite of clustering algorithms that harness textual information as external guidance has emerged (Cai et al., 2023; Li et al., 2024; Qiu et al., 2024; Peng et al., 2025; Li et al., 2025; Jiu et al., 2026; Zhu et al., 2026; Zhong et al., 2026), representing the leading paradigm.

Under this Text-assisted Image Clustering (TIC) paradigm, existing methods generally follow a two-stage CLIP-based (Radford et al., 2021) pipeline: *semantic representation construction* followed by *semantic exploitation*. In Stage 1, they construct dataset-relevant textual semantic representations from large text pools, such as WordNet (Miller, 1995), mainly through heuristic filtering and cross-modal matching. Specifically, Li et al. (2024); Cai et al. (2023); Qiu et al. (2024); Peng et al. (2025); Jiu et al. (2026) rely on WordNet nouns with handcrafted priors, gradient cues, and cross-madal matching, while Zhu et al. (2026); Zhong et al. (2026) further enrich this process with caption-level semantics, optimal transport, or LLM-derived concept-attribute knowledge. In Stage 2, these semantics are injected into clustering through pseudo-labeling, neighborhood consistency, cross-modal distillation, adaptive fusion, or feature enhancement (Cai et al., 2023; Qiu et al., 2024; Li et al., 2024; Jiu et al., 2026; Zhu et al., 2026; Zhong et al., 2026).

Despite their progress, existing TIC methods are constrained by imperfect CLIP image-text alignment (Radford et al., 2021; Singh & Zheng, 2023; Zhang et al., 2024; Abbasi et al., 2025), making

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.

Table 1: Comparison highlighting the regularity and symmetry of S^3 against existing TIC methods.

Method	Aspect	Stage 1	Stage 2
Existing Methods	**Core Design**	Rules/thresholds + P2P matching	Clustering loss
	Solver	**Non-optimized form**	Mini-batch + gradient descent
S^3 **(Ours)**	**Core Design**	**Unified structure-guided objective**	
	Solver	Mini-batch + **coordinate descent**	Mini-batch + **gradient descent**

Stage 1 both fragile and over-engineered. Rule-based filtering (Cai et al., 2023; Qiu et al., 2024; Peng et al., 2025; Jiu et al., 2026; Zhong et al., 2026) depends on heuristic local criteria, and thus often requires dataset-specific thresholds, ratios, or fusion weights. Cross-modal point-to-point matching (Li et al., 2024; Cai et al., 2023; Qiu et al., 2024; Jiu et al., 2026; Zhu et al., 2026) is likewise local: instance-level matching is sensitive to noisy image-text neighbors, whereas center-based matching compresses complex image manifolds into coarse anchors, fails to capture precise semantic boundaries. Ultimately, the semantic representations constructed in Stage 1 remain limited in quality, and prior TIC studies have shown that this can severely degrade final clustering performance (Cai et al., 2023; Li et al., 2024; Qiu et al., 2024; Jiu et al., 2026; Zhu et al., 2026; Zhong et al., 2026).

These observations raise a more fundamental question: *what principle should be used to select textual semantic representations?* We reason about this question as follows: a single image of "a dog on grass" may be locally aligned with both "dog" and "grass", and even with misleading nearby concepts under imperfect cross-modal alignment. The semantic boundary of "dog" becomes reliable only when many dog images are considered together across diverse backgrounds, poses, and appearances; if multiple dog breeds appear, the dataset structure can further shape finer-grained breed-level boundaries. Thus, semantic relevance should not be judged by isolated text candidates, individual image-text pairs, or coarse visual centers alone. Instead, high-quality semantic representations should be selected according to how well they explain the global feature structure of the whole image dataset.

In this paper, we argue that the intrinsic feature **S**tructure of an image dataset plays a pivotal role in **S**haping high-quality textual **S**emantic representations (S^3). Guided by this principle, we advocate replacing instance-level cross-modal matching with alignment between the global feature distribution of the dataset and the textual semantic representations, enabling a more reliable capture of category-level structures and semantic boundaries while mitigating the impact of individual noise. However, this constitutes an NP-hard combinatorial optimization problem. Fortunately, we identify a simple and natural solution: akin to standard network training, we sample mini-batches, compute the clustering loss, and perform *coordinate* descent. With this treatment, our two-stage algorithm surprisingly shares the same objective function and exhibits a clean symmetric structure, as summarized in Table 1. We believe this substantially advances TIC from heuristic semantic construction toward principled optimization. Furthermore, we introduce Fine-Structure guided Optimization (FSO), which leverages image augmentations to elicit finer feature structures, thereby further refining the semantic representations. Extensive experiments validate the effectiveness of our algorithm. Notably, on the challenging Tiny-ImageNet benchmark, we outperform existing methods by 6.0% (ACC).

Our contributions are summarized as follows:

1) We formulate TIC as a unified two-stage optimization problem, where semantic representation construction and clustering network training share the same structure-aware objective.

2) We reveal a symmetric TIC optimization structure: Stage 1 optimizes textual semantics under fixed image features, while Stage 2 optimizes clustering predictions under fixed semantics.

3) We introduce the **S**tructure **S**hapes **S**emantics (S^3) principle and FSO, which derives textual semantics from the global feature structure of unlabeled data.

4) Our method achieves state-of-the-art performance on seven benchmarks with few dataset-specific hyperparameters.

2 Related Work

2.1 Vision-Language Models

Vision-Language Models (VLMs) learn cross-modal correspondences between images and texts through large-scale pre-training, enabling zero-shot transfer to diverse downstream tasks. Contrastive learning-based models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) align visual and textual representations in a shared embedding space, with CLIP demonstrating strong

generalization via a symmetric InfoNCE loss and ALIGN showing robustness to noisy image-text pairs. Subsequent works have advanced contrastive pretraining through data-efficient strategies (Li et al., 2021a), region-word alignment (Yao et al., 2021), and multi-level semantic modeling (Gao et al., 2022). Generative approaches like COCA (Yu et al., 2022) and FLAVA (Singh et al., 2022) capture richer contextual information via masked modeling or image-to-text generation, complementing contrastive objectives. Alignment-based methods such as GLIP (Li et al., 2022b) and FIBER (Dou et al., 2022) explicitly model image-text matching, extending to region-level alignment for dense prediction tasks. In this paper, we focus on leveraging cross-modal alignment to enhance image clustering.

2.2 Image Clustering

Image Clustering (IC) partitions unlabeled images into semantically meaningful groups. Classical clustering (MacQueen et al., 1967; Ester et al., 1996; Zelnik-Manor & Perona, 2005), deep clustering (Yang et al., 2016; Jiang et al., 2016; Mukherjee et al., 2019), and recent self-supervised methods (Ji et al., 2019; Li et al., 2021b; Van Gansbeke et al., 2020; Li et al., 2022c) have progressively improved visual representations and assignments, but relying solely on internal visual information still limits performance. Text-assisted Image Clustering (TIC) introduces textual semantics through VLMs (Radford et al., 2021; Singh & Zheng, 2023) for external guidance. Existing TIC methods generally first filter relevant semantic representations from a large textual candidate set, and then use the selected semantics to assist second-stage clustering or training. Early methods such as SIC (Cai et al., 2023), TAC (Li et al., 2024), MCA (Qiu et al., 2024), and GradNorm (Peng et al., 2025) mainly select WordNet nouns (Miller, 1995) through cross-modal matching, handcrafted priors, or gradient clues, while recent methods enrich this process with caption-level semantics (CAE (Zhu et al., 2026)), selective spatial text (SATC (Jiu et al., 2026)), and LLM-derived concept-attribute knowledge (KEC (Zhong et al., 2026)). However, semantic filtering is easily affected by imperfect image-text alignment, and existing methods adopt largely inconsistent mechanisms across two stages, leading to many hyperparameters and complex pipelines. In contrast, we reformulate semantic representation construction as an optimization problem, thereby successfully unifying the two stages. This formulation greatly reduces the number of hyperparameters and simplifies the TIC pipeline.

3 Preliminary

3.1 Problem Formulation

Notations. In this work, we focus on addressing Image Clustering (IC) using Vision-Language Models (VLMs), a paradigm commonly referred to as Text-assisted Image Clustering (TIC) (Li et al., 2024). Following prior studies (Li et al., 2024; Qiu et al., 2024; Cai et al., 2023; Li et al., 2025), we adopt the CLIP model (Radford et al., 2021). Given an unlabeled dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ and a desired number of clusters K, the objective of TIC is to leverage CLIP's multimodal capabilities to supplement image datasets with discriminative cues from the textual modality, thereby accurately grouping images into semantically meaningful clusters. CLIP consists of an image encoder f and a text encoder g. Given a set of text descriptions $\mathcal{T} = \{t_1, t_2, \ldots, t_K\}$ and an image $\mathbf{x}_i$, CLIP produces image embeddings $f(\mathbf{x}_i) \in \mathbb{R}^D$ and text embeddings $g(t_j) \in \mathbb{R}^D, \forall j \in \{1, 2, \ldots, K\}$. Both embeddings are L_2-normalized. Subsequently, zero-shot classification can be performed, which we denote as:

$$\mathbf{p}_i = h(\mathbf{x}_i; \mathcal{T}) = [p_1, p_2, \ldots, p_K]^T, \tag{1}$$

$$p_j = \frac{\exp(g(t_j)^T f(\mathbf{x}_i)/\tau)}{\sum_{k=1}^K \exp(g(t_k)^T f(\mathbf{x}_i)/\tau)}, \tag{2}$$

where $\tau = 0.01$ is the softmax temperature. In addition, following prior works (Li et al., 2024; Cai et al., 2023; Qiu et al., 2024), WordNet (Miller, 1995) nouns are used to obtain semantic representation candidates, denoted as $\mathcal{T}^{all}$.

3.2 Heuristic Strategies for Semantic Matching

Existing TIC (Li et al., 2024; Qiu et al., 2024; Cai et al., 2023; Peng et al., 2025; Jiu et al., 2026) methods typically first employ a matching (or filtering) algorithm to select a subset of nouns $\hat{\mathcal{T}} \subset \mathcal{T}^{all}$ most relevant to $\mathcal{D}$, which we refer to as instance-level or category-level **semantic representations**.

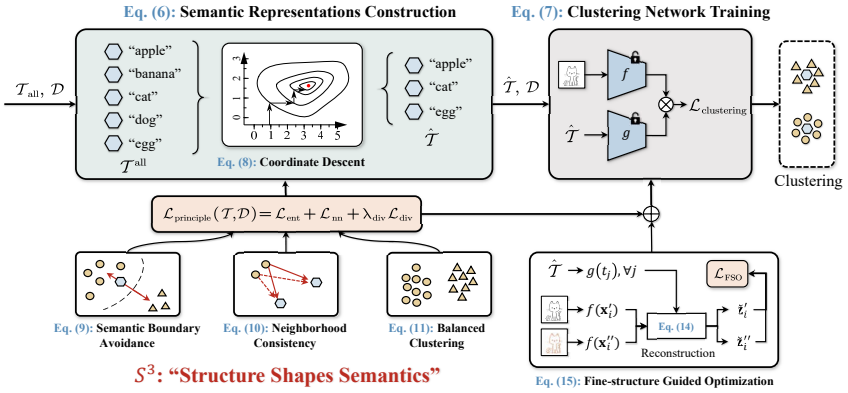


Figure 1: **Framework of our method.** Our approach is built upon a principled, globally scoped optimization framework that enhances the quality of textual semantic representations and thereby strengthens clustering performance. It consists of two sequential modules—*semantic representation construction* ( ) and *clustering network training* ( ). These two modules respectively rely on semantic-quality criteria and the clustering loss, while sharing a unified set of core components ( ) within our framework. All components are guided by the underlying principle that "Structure Shapes Semantics" (S^3).

In this work, we summarize the matching (or filtering) strategies of Li et al. (2024); Qiu et al. (2024); Cai et al. (2023); Jiu et al. (2026) in a formalized manner as follows:

1) Uniqueness (Cai et al., 2023): This strategy aims to filter out words that convey overly general meanings, as:

$$\hat{\mathcal{T}} = \{t_i \in \mathcal{T}^{all} | 1 - \text{sim}(g(t_i), \mathbf{e}) > \text{thershold}\}, \tag{3}$$

where $\mathbf{e}$ is the center of all text embeddings of $\mathcal{T}^{all}$.

2) Hierarchy (Qiu et al., 2024): Based on the hierarchical structure of WordNet (Miller, 1995), this strategy selects lower-level nouns, which are assumed to contain finer-grained information, as:

$$\hat{\mathcal{T}} = \{t_i \in \mathcal{T}^{all} | \text{ Level}(t_i) < \text{thershold}\}. \tag{4}$$

3) Point-to-point (P2P) Cross-modal Matching (Cai et al., 2023; Qiu et al., 2024; Li et al., 2024; Jiu et al., 2026): This strategy leverages image-text alignment as the core mechanism, using queries $\mathbf{u} \in \mathcal{U}$ to select semantically relevant nouns from the candidate set, as:

$$\hat{\mathcal{T}} = \{t_i | t_i = \text{argmax}_{t_i \in \mathcal{T}^{all}} \, g(t_i)^T \mathbf{u}, \forall \mathbf{u} \in \mathcal{U}\}. \tag{5}$$

When $\mathbf{u}$ are image embeddings, Eq. (5) defines cross-modal nearest neighbors (Cai et al., 2023; Li et al., 2024; Qiu et al., 2024); when $\mathbf{u}$ are K-means centroids, it defines center-based mapping (Cai et al., 2023; Li et al., 2024; Jiu et al., 2026).

Using these strategies, individually or in combination, existing methods (Li et al., 2024; Qiu et al., 2024; Cai et al., 2023) generate instance- or category-level semantic representations to enrich discriminative signals. However, these strategies are largely heuristic and constrained by imperfect cross-modal alignment (Radford et al., 2021; Singh & Zheng, 2023; Zhang et al., 2024; Abbasi et al., 2025). We discuss their limitations in detail in the Section C. Given the strong correlation between semantic representation quality and clustering performance (Qiu et al., 2024; Cai et al., 2023; Li et al., 2024), existing TIC methods often yield suboptimal results.

4 METHOD

In this paper, we focus on leveraging the global feature structure of image datasets to shape high-quality textual semantic representations for boosting clustering performance. In Section 4.1, we formalize the semantic representation optimization and introduce a simple optimization algorithm. In Section 4.2, we present the principle of "Structure Shapes Semantics" (S^3) along with its corresponding computable objective. In Section 4.3, we further propose leveraging finer-structure to guide the refinement of semantic representations. An overview of the entire pipeline is provided in Fig. 1. Additionally, a detailed description of our algorithmic pipeline can be found in the supplementary material.

4.1 SEMANTIC REPRESENTATION OPTIMIZATION

For a systematic study, we first provide a formal definition of the mainstream TIC (Li et al., 2024; Qiu et al., 2024; Cai et al., 2023) framework as follows:

1) Semantic Representations Construction. Construct a set of textual semantic representations for $\mathcal{D}$ following the principle:

$$\hat{\mathcal{T}} = \operatorname{argmin}_{\mathcal{T} \subset \mathcal{T}^{all}} \ \mathcal{L}_{\text{principle}}(\mathcal{T}, \mathcal{D}). \tag{6}$$

2) Clustering Network Training. Optimize the clustering network parameterized by θ using $\hat{\mathcal{T}}$ and $\mathcal{D}$:

$$\min_{\theta} \ \mathcal{L}_{\text{clustering}}(\hat{\mathcal{T}}, \mathcal{D}; \theta), \tag{7}$$

where $\mathcal{L}_{\text{principle}}$ and $\mathcal{L}_{\text{clustering}}$ are left undefined to ensure generality.

Based on the formal definition, the heuristic strategies in Section 3.2 can be summarized as two formulations: $\mathcal{L}_{\text{principle}}(t_i)$ and $\mathcal{L}_{\text{principle}}(\mathcal{T}, \mathbf{u})$. The former corresponds to *uniqueness* and *hierarchy*, while the latter corresponds to *P2P cross-modal matching*. From these formulations, it becomes evident that such heuristic strategies do not construct the semantic representation set $\hat{\mathcal{T}}$ from a global perspective based on the entire dataset $\mathcal{D}$. While they make Eq. (6) easy to optimize, they struggle to handle noise introduced by imperfect cross-modal alignment (Radford et al., 2021) and noisy semantic space, failing to accurately capture category boundaries.

In this work, we adhere strictly to the formulation in Eq. (6), aiming to derive K category-level semantic representations $\hat{\mathcal{T}} = \{t_1, t_2, \ldots, t_K\} \subset \mathcal{T}^{all}$ from a global perspective. This leads to a combinatorial optimization problem—selecting K non-overlapping elements from $\mathcal{T}^{all}$ to minimize $\mathcal{L}_{\text{principle}}$ on $\mathcal{D}$—which inevitably suffers from combinatorial explosion. To circumvent this, we introduce a simple yet natural optimization approach: akin to neural network training, we sample a mini-batch $\mathcal{B} \subset \mathcal{D}$, calculate $\mathcal{L}_{\text{principle}}(\mathcal{T}, \mathcal{B})$, and iteratively optimize $\mathcal{T}$ via *coordinate* descent. We formalize one iteration of *cyclic coordinate descent* as:

$$\begin{aligned} t_j^{(s+1)} &= \operatorname*{argmin}_{t_j \in \mathcal{T}^{all} \cup \{t_j^{(s)}\}} \mathcal{L}_{\text{principle}}(\mathcal{T}', \mathcal{B}_j^{(s)}), \quad \forall\, j \\ \mathcal{T}' &= \{t_1^{(s+1)}, \ldots, t_{j-1}^{(s+1)}, t_j, t_{j+1}^{(s)}, \ldots, t_K^{(s)}\}, \end{aligned} \tag{8}$$

where s is the iteration step and $\mathcal{B}_j^{(s)} \subset \mathcal{D}$. Note that $t_j^{(s)}$ is updated sequentially and $\mathcal{T}$ is randomly initialized.

Empirical results demonstrate the effectiveness and convergence of our optimization algorithm, and a formal convergence analysis is provided in Section B. To maintain efficiency without sacrificing performance, we sample N_{text} nouns from $\mathcal{T}^{all}$ at each step. The image batchsize is set to N_{image}, and the optimization runs for N_{iter} iterations. The computational complexity of Eqn. 8 is $O(N_{\text{text}} N_{\text{image}} N_{\text{iter}} K)$, where the first three factors are constants, and the last term indicates linear growth with respect to K.

4.2 STRUCTURE SHAPES SEMANTICS

Although we have established a global framework for semantic representation optimization, the principles for deriving high-quality semantic representations have yet to be specified. As noted in the introduction, we posit that the intrinsic feature **S**tructure of the dataset plays a pivotal role in **S**haping high-quality **S**emantic representations (S^3). Accordingly, we propose the following key considerations and instantiate our S^3 principle into a computable metric for category-level semantic representation quality.

1) Semantic boundary avoidance. High-quality representations should lie far from semantic boundaries. We evaluate the prediction entropy for each sample via an entropy-minimization loss (Grandvalet & Bengio, 2004; Elsayed et al., 2018), which implicitly captures the class feature distribution, as:

$$\mathcal{L}_{ent}(\mathbf{x}_i) = -\sum_{k=1}^{K} p_k \log p_k. \tag{9}$$

Table 2: Clustering performance on seven widely used image clustering datasets. The best and second-best results are highlighted in **bold** and underlined, respectively. Higher values indicate better performance (↑). Δ represents the improvement of our method over the second-best result (green for >3.0% improvement).

Dataset	STL10			CIFAR10			CIFAR20			ImageNet-Dogs			Tiny-ImageNet			DTD			UCF-101		
Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
K-means (MacQueen et al., 1967)	19.2	12.5	6.1	22.9	8.7	4.9	13.0	8.4	2.8	10.5	5.5	2.0	2.5	6.5	0.5	–	–	–	–	–	–
CC (Li et al., 2021b)	85.0	76.4	72.6	79.0	70.5	63.7	42.9	43.1	26.6	42.9	44.5	27.4	14.0	34.0	7.1	–	–	–	–	–	–
GCC (Zhong et al., 2021)	78.8	68.4	63.1	85.6	76.4	72.8	47.2	47.2	30.5	52.6	49.0	36.2	13.8	34.7	7.5	–	–	–	–	–	–
TCC (Shen et al., 2021)	81.4	73.2	68.9	90.6	79.0	73.3	49.1	47.9	31.2	59.5	55.4	41.7	–	–	–	–	–	–	–	–	–
SCAN (Van Gansbeke et al., 2020)	80.9	69.8	64.6	88.3	79.7	77.2	50.7	48.6	33.3	59.3	61.2	45.7	42.0	69.8	33.2	46.4	59.4	31.7	61.1	79.7	53.1
TCL (Li et al., 2022c)	86.8	79.9	75.7	88.7	81.9	78.0	53.1	52.9	35.7	64.4	62.3	51.6	–	–	–	–	–	–	–	–	–
DPAC (Yan et al., 2024)	92.6	84.5	84.7	90.7	82.7	81.2	51.6	52.9	36.2	65.5	60.2	50.0	–	–	–	–	–	–	–	–	–
CDC (Jia et al., 2025)	93.0	85.8	85.6	94.9	89.3	89.5	61.7	60.9	46.6	79.2	76.5	70.0	33.9	47.5	19.9	–	–	–	–	–	–
CLIP (K-means) (Radford et al., 2021)	94.3	91.7	89.1	74.2	70.3	61.6	45.5	49.9	28.3	38.1	39.8	20.1	55.2	77.6	45.0	42.6	57.3	27.4	58.2	79.5	47.6
SIC (Cai et al., 2023)	98.1	95.4	95.9	92.7	84.8	84.6	58.4	59.3	44.0	71.3	71.8	58.6	61.2	80.5	51.4	45.9	59.6	30.5	61.9	81.0	53.6
TAC (Li et al., 2024)	98.2	95.5	96.1	91.9	83.3	83.1	60.7	61.1	44.8	83.0	80.6	72.2	–	–	–	50.1	62.1	34.4	**68.7**	82.3	**60.1**
MCA (Qiu et al., 2024)	98.2	95.5	96.0	92.8	85.0	84.9	61.2	60.6	45.5	77.9	75.1	64.3	61.9	81.1	52.3	–	–	–	–	–	–
KEC (Zhong et al., 2026)	97.9	95.1	95.5	90.7	81.9	80.6	51.8	66.4	37.3	-	-	-	-	-	-	47.4	60.7	31.5	62.5	81.9	53.2
CAE (Zhu et al., 2026)	98.2	95.5	96.7	90.9	81.7	80.5	60.9	62.8	45.9	77.5	77.9	66.3	-	-	-	46.7	62.1	31.9	63.5	82.8	54.9
GradNorm (Peng et al., 2025)	98.3	95.6	96.2	91.1	82.6	81.5	60.6	61.3	43.6	81.2	81.0	70.9	–	–	–	50.9	63.1	34.2	62.7	82.9	53.2
S^3 (Ours)	**98.6**	**96.4**	**96.9**	**96.6**	**92.3**	**92.8**	**62.8**	**65.6**	**49.6**	**86.0**	**86.5**	**79.8**	**67.9**	**82.6**	**57.7**	**50.6**	**63.9**	**37.4**	68.3	**84.3**	59.4
Δ	+0.3	+0.8	+0.2	+1.7	**+3.0**	**+3.3**	+1.1	+2.8	**+3.0**	**+3.0**	**+5.5**	**+7.6**	**+6.0**	+1.5	**+5.4**	+0.5	+0.8	**+3.0**	-0.4	+1.4	-0.7

2) Neighborhood consistency. High-quality representations should yield similar predictions for similar images (Belkin & Niyogi, 2003; Belkin et al., 2006). By incorporating neighborhood relations, the structure of the feature space is explicitly encoded into the metric, as:

$$\mathcal{L}_{nn}(\mathbf{x}_i) = -\mathbf{p}_i^T \mathbf{p}_i - \sum_{\mathbf{x_j} \in \mathcal{N}_i} \mathbf{p}_j^T \mathbf{p}_i \ , \tag{10}$$

where $\mathcal{N}_i$ contains N_{nn} nearest neighbors of $\mathbf{x}_i$.

3) Balanced clustering. High-quality semantic representations should ensure balanced structures across clusters:

$$\mathcal{L}_{div}(\mathbf{x}_i) = \sum_{k=1}^{K} \bar{p}_k \log \bar{p}_k, \tag{11}$$

where $\bar{\mathbf{p}} = [\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_K] = \mathbb{E}_{\mathbf{x}_i \in \mathcal{D}}\ \mathbf{p}_i$.

Overall objectives. Finally, we instantiate our $L_{\text{principle}}$ as a computable metric:

$$\mathcal{L}_{S^3} = \mathcal{L}_{ent} + \mathcal{L}_{nn} + \lambda_{div}\mathcal{L}_{div}. \tag{12}$$

It is worth emphasizing that in the TIC framework, the proposed $\mathcal{L}_{S^3}$ serves as a feature structure guided criterion for *deriving the optimal semantic representations* $\hat{\mathcal{T}}$. The counterpart of $\mathcal{L}_{S^3}$ in existing methods is essentially the heuristic strategy (Eq. (6)), rather than the clustering loss used for network training (Eq. (7)). We advance prior handcrafted semantic matching strategies into a coherent optimization framework grounded in formal principles.

Homogeneity of objectives. Furthermore, it is evident that $\mathcal{L}_{S^3}$ can also serve as the $\mathcal{L}_{\text{clustering}}$. Accordingly, in our complete algorithm, we have:

$$\mathcal{L}_{\text{principle}} = \mathcal{L}_{\text{clustering}} = \mathcal{L}_{S^3}. \tag{13}$$

Notably, our Eq. (13) reveals the intrinsic homogeneity between textual semantic representation construction and clustering network training within the TIC framework.

4.3 Fine-Structure Guided Optimization

We have established an optimization approach for Eq. (6) and instantiated a computable metric following the S^3 principle. However, when the images contain numerous visually similar categories, the aforementioned coarse feature structure may be insufficient to shape precise semantic representations. To further enhance the quality of semantic representations, we turn to finer-grained feature structures.

Specifically, for an augmented view $\mathbf{x}'_i$ of an image, we first compute its coordinates in the semantic representation space: $\mathbf{p}'_i = h(\mathbf{x}'_i; \hat{\mathcal{T}})$. Then, we use the semantic representations as basis vectors to reconstruct the image features:

$$\tilde{\mathbf{z}}'_i = [g(t_1), g(t_2), \ldots, g(t_K)]\ \mathbf{p}'_i. \tag{14}$$

The reconstructed semantically meaningful features are expected to be consistent across different augmentations, which is encouraged using a contrastive objective:

$$\mathcal{L}_{FSO} = -\log \frac{l(\tilde{\mathbf{z}}'_i, \tilde{\mathbf{z}}''_i)}{l(\tilde{\mathbf{z}}'_i, \tilde{\mathbf{z}}''_i) + \sum_{\tilde{\mathbf{z}} \in \mathcal{M}} l(\tilde{\mathbf{z}}'_i, \tilde{\mathbf{z}})}, \tag{15}$$

where $\tilde{\mathbf{z}}'_i$ and $\tilde{\mathbf{z}}''_i$ are the reconstructed features of two different augmented views of image $\mathbf{x}_i$, $\mathcal{M}$ includes reconstructed features of all other samples in the mini-batch, and $l(\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2) = \exp(\cos(\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2)/\tau_{co})$. Unlike conventional feature space (Chen et al., 2020b;a; He et al., 2020) or image-text contrastive learning (Radford et al., 2021), our $\mathcal{L}_{FSO}$ directly optimizes textual category-level semantic representations, with fine feature structures injected via image augmentation. During training, $\mathcal{L}_{FSO}$ is incorporated into the $\mathcal{L}_{\text{clustering}}$ (Eq. (7)).

5 Experiments

5.1 Experimental Setup

Datasets. For a fair comparison, we adopt the conventional experimental setup of the Text-assisted Image Clustering (Li et al., 2024; Cai et al., 2023) methods and evaluate our algorithm on seven public datasets. The evaluation covers simple datasets such as STL10 (Coates et al., 2011), CIFAR10 (Krizhevsky & Hinton, 2009), CIFAR20 (Krizhevsky & Hinton, 2009), and ImageNet-Dogs (Chang et al., 2017); datasets with a larger number of categories such as DTD (Cimpoi et al., 2014) and UCF101 (Soomro et al., 2012); the most challenging Tiny-ImageNet (Deng et al., 2009), which involves a large number of images and 200 categories.

Evaluation metrics. We follow the same evaluation protocol as existing studies (Li et al., 2024; Cai et al., 2023). Specifically, clustering performance is assessed using Accuracy (ACC), Normalized Mutual Information (NMI) (McDaid et al., 2011), and Adjusted Rand Index (ARI) (Hubert & Arabie, 1985), where ACC is computed by performing Hungarian matching between cluster assignments and ground-truth labels.

Implementation details. Following existing studies, we adopt the large-scale pretrained CLIP (Radford et al., 2021) model, using a Transformer (Vaswani et al., 2017) and ViT-B/32 (Dosovitskiy et al., 2020) as the text and image encoders, respectively. All hyperparameters in our algorithm are listed as follows: $\tau_{co} = 0.3, \lambda_{div} = 2.0$. Except for setting $\tau_{co} = 0.2$ on UCF101 and Tiny-ImageNet due to a larger number of clusters, no dataset-specific hyperparameter tuning is performed. Sampling and iteration parameters are as follows: $N_{\text{text}} = 5000, N_{\text{image}} = 2000, N_{\text{iter}} = 3$. Empirically, they empirically correlate positively with performance. To preserve pretrained knowledge and ensure robust clustering, we apply LoRA (Hu et al., 2021; Smith et al., 2023) fine-tuning during clustering network training. The rank is set to 8, and LoRA is applied only to the query (W_q) and value (W_v) weights of the self-attention modules. During clustering network training, we adopt the SGD optimizer with a learning rate of 0.0005 and cosine annealing scheduling (Loshchilov & Hutter, 2016).

5.2 Main Results

As shown in Table 2, our method consistently achieves the best or second-best performance across seven mainstream image clustering benchmarks, demonstrating remarkable generalization and robustness. On standard image classification datasets such as STL10 and CIFAR20, our S^3 maintains stable superiority. On DTD and UCF101, it performs on par with TAC. Notably, on CIFAR10 and ImageNet-Dogs, S^3 surpasses the current state-of-the-art methods by 3.8% and 3.0% in ACC, respectively. Particularly, on the most challenging dataset—Tiny-ImageNet with the largest number of categories—our method outperforms all competitors by a substantial margin of 6.0% ACC. We attribute this advantage to the high-quality textual semantic representations optimized by our framework, where leveraging global feature structure becomes increasingly beneficial for handling imperfect cross-modal alignment as dataset complexity grows.

5.3 Analysis

Effectiveness of principled optimization. In this work, we aim to identify an optimal set of category-level semantic representations to facilitate image clustering. In prior work, SIC (Cai et al., 2023) constructs category-level semantic representations based on uniqueness and point-to-point cross-modal matching. To validate the superiority of our method, we evaluate the clustering performance

Table 3: Quantitative study of semantic representation quality. The category-level semantic representations obtained by each method are evaluated for clustering performance before training the clustering network.

Method	CIFAR20			ImageNet-Dogs			Tiny-ImageNet		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
SIC	25.3	30.6	7.9	23.9	32.9	9.2	21.4	58.3	14.9
S^3 (Ours)	**50.2**	**50.8**	**33.9**	**72.7**	**73.4**	**59.9**	**65.8**	**81.3**	**54.3**

Table 4: Comparison of clustering performance between directly using raw image features and our reconstructed semantically meaningful features in contrastive loss.

Reconstructed	CIFAR20			ImageNet-Dogs			Tiny-ImageNet		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
✗	61.9	62.4	46.8	80.3	81.6	70.4	66.8	82.3	56.4
✓	**62.8**	**65.6**	**49.6**	**86.0**	**86.5**	**79.8**	**67.9**	**82.6**	**57.7**
Δ	+0.9	+3.2	+2.8	+5.7	+4.8	+9.4	+1.1	+0.3	+1.3

of the semantic representations from both S^3 and SIC prior to clustering network training. As shown in Table 3, our algorithm demonstrates substantial gains across all three datasets.

Table 5: Evaluation of the impact of global feature structure completeness on semantic optimization. Experiments are performed on Tiny-ImageNet by varying the dataset sampling ratio.

Metric	Sampling Ratio of $\mathcal{D}$			
	1%	5%	20%	100%
ACC	50.31	60.76	64.40	**65.80**
NMI	73.66	78.98	80.89	**81.30**
ARI	39.82	49.69	53.18	**54.30**

Table 6: Clustering performance of proposed S^3 with different combinations of the loss functions. "*" indicates that the loss is used only in Eq. (6) and not in Eq. (7).

	Components				CIFAR20			ImageNet-Dogs			Tiny-ImageNet		
	$\mathcal{L}_{ent}$	$\mathcal{L}_{nn}$	$\mathcal{L}_{div}$	$\mathcal{L}_{FSO}$	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
1	*	*	*	✓	55.1	54.3	38.8	73.0	71.0	59.6	66.3	81.4	54.4
2		✓	✓	✓	60.2	63.0	46.7	81.5	80.9	72.7	66.2	82.1	56.3
3	✓		✓	✓	59.3	62.9	46.1	74.8	74.9	62.9	63.9	79.9	51.9
4	✓	✓		✓	36.8	42.6	10.5	62.7	69.2	44.9	61.9	82.1	50.5
5	✓	✓	✓		60.1	61.6	45.7	81.9	81.0	71.7	67.0	**82.8**	57.3
6	✓	✓	✓	✓	**62.8**	**65.6**	**49.6**	**86.0**	**86.5**	**79.8**	**67.9**	82.6	**57.7**

Role of global feature structure. Here, we quantitatively investigate the role of global feature structure in our method. To isolate the influence of the feature structure during the optimization process, we modulate the completeness of the structural information by sampling the training dataset $\mathcal{D}$. Specifically, we sample subsets of $\mathcal{D}$ with ratios of $\{1\%, 5\%, 20\%, 100\%\}$ to participate in the optimization of the objective function (Eq. (6)). For instance, a ratio of 1% indicates that only 1% of the samples from the original dataset $\mathcal{D}$ are utilized in the first stage. We report the performance of the optimized semantic set $\hat{\mathcal{T}}$ prior to clustering network training. As shown in the Table 5, clustering performance exhibits a monotonic increase with the sampling ratio and tends to saturate when the data volume is sufficiently large. These results support the necessity of global optimization. Moreover, these results indicate that when the sample size is limited, the partial intrinsic feature structure struggles to mitigate the impact of noise. Therefore, we conclude that the global feature structure of the samples plays a pivotal role in shaping high-quality semantic representations, when image-text alignment is imperfect and substantial noise is present.

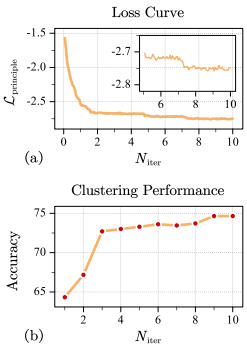


Figure 2: Curves of the loss and clustering accuracy over coordinate-descent iterations on the ImageNet-Dogs dataset. The inset in panel (a) shows a magnified local region.

Effectiveness of fine-structure guided optimization. In Section 4.3, we introduced a Fine-Structure guided Optimization (FSO) strategy that reconstructs augmented image features using category-level semantic representations and applies a contrastive loss to refine these semantics. Here, we conduct an additional experiment where the features used for contrastive learning are taken directly from the image embeddings, to demonstrate that the performance gain of FSO does not stem from conventional self-supervised contrastive learning, but rather from optimizing the semantic representations themselves. As shown in Table 4, contrastive learning on the reconstructed, semantically meaningful features consistently outperforms that on raw features across all three datasets, with particularly notable improvement on the fine-grained ImageNet-Dogs dataset—confirming our hypothesis.

Quantitative study on coordinate descent convergence. In this study, we quantitatively examine the convergence of Eq. (8). As shown in Fig. 2, we set $N_{\text{iter}} = 10$ and track the evolution of $\mathcal{L}_{\text{principle}}$ and clustering accuracy. Note that the loss is computed at each step corresponding to the update of a specific t_j, and N_{iter} is incremented only after all t_j have been updated once. Fig. 2 (a) demonstrates that the loss is effectively reduced and gradually stabilizes as iterations progress. Fig. 2 (b) shows that, for sufficiently large N_{iter}, clustering performance during the coordinate descent stage also approaches saturation. Notably, subsequent network training will further enhance clustering results.

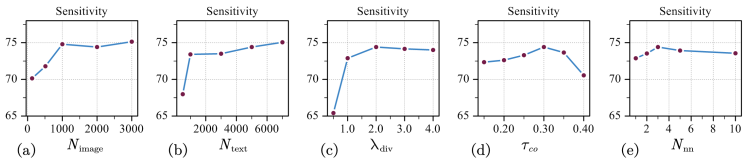


Figure 3: Accuracy (%) curves with respect to varying hyperparameters. Results are averaged over CIFAR20 and ImageNet-Dogs. (a–e) correspond to N_{image}, N_{text}, λ_{div}, τ_{co}, and N_{nn}, respectively.

Loss ablation studies. As shown in Table 6, we conduct ablation studies on the different loss components used in our method ($\mathcal{L}_{ent}$, $\mathcal{L}_{nn}$, $\mathcal{L}_{div}$, and $\mathcal{L}_{FSO}$). Note that in our full algorithm, only the first three losses are applied during the stage corresponding to Eq. (6), while all losses are employed during the clustering network training (Eq. (7)). Losses denoted by * in Table 6 are only employed in Eq. (6). If a loss is excluded, it is omitted in both Eqs. (6) and (7). Experiments are conducted on standard (CIFAR20), fine-grained (ImageNet-Dogs), and the most challenging dataset (Tiny-ImageNet) to thoroughly assess the contribution of each component. As indicated in Table 6, our full method (#6) substantially outperforms alternative combinations (#1–5), demonstrating that integrating "semantic boundary avoidance", "neighborhood consistency", "balanced clustering", and fine-structure yields the most accurate high-quality semantic representations. Additionally, focusing on the role of $\mathcal{L}_{FSO}$, #1 shows that using $\mathcal{L}_{FSO}$ alone already achieves a significant improvement over the state-of-the-art MCA (Qiu et al., 2024) on Tiny-ImageNet, confirming the importance of fine-structure for semantic representation optimization. Furthermore, comparing #5 with #6, $\mathcal{L}_{FSO}$ also substantially compensates for other losses on the fine-grained ImageNet-Dogs dataset.

Basset
Blenheim spaniel
Brittany spaniel
Chow
Clumber
Doberman
Giant schnauzer
Golden retriever
Groenendael
Kelpie
Maltese dog
Norwegian elkhound
Pug
Shetland sheepdog
Welsh springer spaniel

(a) GT Class Names

Basenji
Basset
Belgian Sheepdog
Brittany Spaniel
Chowchow
Doberman
Dog Show
Giant Schnauzer
Golden Retriever
Kelpie
King Charles Spaniel
Maltese Dog
Norwegian Elkhound
Pug
Shetland Sheepdog

(b) Optimized Nouns

Figure 4: Optimized $\hat{\mathcal{T}}$ on ImageNet-dogs. The figure shows the category-level textual semantic representations obtained by our Eq. (6).

Sensitivity analysis of hyperparameters. As shown in Fig. 3, we evaluate the sensitivity of our algorithm to hyperparameters on CIFAR20 and ImageNet-Dogs. Firstly, the parameters N_{image} and N_{text} determine the number of sampled images and texts per step in the coordinate descent optimization (Eq. (8)). As illustrated in Fig. 3 (a–b), the clustering accuracy stabilizes once these values exceed a certain threshold. Within computational limits, increasing N_{image} and N_{text} can further improve performance. In addition, Fig. 2 indicates that N_{iter} behaves similarly. Meanwhile, both λ_{div} and τ_{co} exhibit stable behavior around their default values (Fig. 3 (c-d)).

Illustration of optimized $\hat{\mathcal{T}}$. To illustrate the effectiveness of our optimization framework, we compare the optimized semantic set $\hat{\mathcal{T}}$ with the ground-truth class names. As shown in Fig. 4, our method successfully recovers nearly all target categories with high lexical or semantic fidelity. Notably, even when the optimized results do not match the ground-truth names exactly, they remain semantically coherent and visually grounded. For instance, Basenji emerges as a plausible alternative within a similar dog-breed subspace; Belgian Sheepdog aligns closely with Groenendael, its actual breed group. These findings demonstrate that our method gracefully handles subtle naming variations and hierarchical relationships between visually similar breeds. This strongly supports our claim that feature-structure–driven optimization yields robust, high-fidelity textual semantics even under imperfect cross-modal alignment.

6 Conclusion

This work reconceptualizes Text-assisted Image Clustering (TIC) through the **S**tructure **S**hapes **S**emantics (S^3) principle. Departing from heuristic priors, we reformulate semantic representation construction as a global optimization problem guided by the intrinsic dataset structure. We propose a coordinate descent algorithm utilizing clustering loss for semantic selection, complemented by Fine-Structure Guided Optimization (FSO) for further refinement. Extensive experiments across seven benchmarks demonstrate that our framework consistently achieves state-of-the-art performance, effectively advancing TIC from empirical heuristics toward a principled optimization paradigm.

AI USE STATEMENT

In the article, we only used LLMs to polish our writing and did not use them for any other assistance.

REFERENCES

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Analyzing clip's performance limitations in multi-object scenarios: A controlled high-resolution study. *arXiv preprint arXiv:2502.19828*, 2025.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11), 2006.

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 *International Conference on Computer Vision*, pp. 5879–5887, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

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

Adam Coates and Andrew Y Ng. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade: Second Edition*, pp. 561–580. Springer, 2012.

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*, pp. 215–223, 2011.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022.

Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. *Advances in neural information processing systems*, 31, 2018.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pp. 226–231, 1996.

Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022.

K Chidananda Gowda and G Krishna. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition*, 10(2):105–112, 1978.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

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*, pp. 9865–9874, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Yuheng Jia, Jianhong Cheng, Hui LIU, and Junhui Hou. Towards calibrated deep clustering network. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=JvH4jDDcG3.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*, 2016.

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

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

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*, pp. 12888–12900. PMLR, 2022a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022b.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021a.

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, pp. 8547–8555, 2021b.

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, 2022c.

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

Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, and Zhaohua Yang. Deep adversarial multi-view clustering network. In *IJCAI*, pp. 2952–2958, 2019.

Zihan Li, Wei Sun, Jing Hu, Jianhua Yin, Jianlong Wu, and Liqiang Nie. Self-enhanced image clustering with cross-modal semantic consistency. *arXiv preprint arXiv:2508.01254*, 2025.

Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19325–19337, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 281–297. Oakland, CA, USA, 1967.

Aaron F McDaid, Derek Greene, and Neil Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*, 2011.

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

Nairouz Mrabah, Mohamed Bouguessa, and Riadh Ksantini. Adversarial deep embedded clustering: on a better trade-off between feature randomness and feature drift. *KDE*, 2020.

Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan. Clustergan: Latent space clustering in generative adversarial networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 4610–4617, 2019.

Feiping Nie, Zinan Zeng, Ivor W Tsang, Dong Xu, and Changshui Zhang. Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. *IEEE Transactions on Neural Networks*, 22 (11):1796–1808, 2011.

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.

Vignesh Prasad, Dipanjan Das, and Brojeshwar Bhowmick. Variational clustering: Leveraging variational autoencoders for image clustering. *arXiv preprint arXiv:2005.04613*, 2020.

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, pp. 14695–14703, 2024.

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.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15638–15650, 2022.

Jaskirat Singh and Liang Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. *Advances in Neural Information Processing Systems*, 36:70799–70811, 2023.

James Seale Smith, Paola Cascante-Bonilla, Assaf Arbelle, Donghyun Kim, Rameswar Panda, David Cox, Diyi Yang, Zsolt Kira, Rogerio Feris, and Leonid Karlinsky. Construct-vl: Data-free continual structured vl concepts learning. In *CVPR*, 2023.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5147–5156, 2016.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, pp. 1601–1608, 2005.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pp. 310–325. Springer, 2024.

Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In *European Conference on Computer Vision*, pp. 41–58. Springer, 2024.

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*, pp. 9224–9233, 2021.

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

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.

A Algorithm Overview

Algorithm 1 rigorously outlines the workflow of our method. Similar to existing TIC approaches, it consists of two sequential stages: semantic representation construction and clustering network training. However, unlike prior works, we formalize the first stage as an optimization problem, rather than relying on heuristic strategies. We design principled objective functions grounded in our proposed principle — Structure Shapes Semantics — and solve them via a coordinate-descent-based optimization algorithm.

Notably, the objective functions employed in the two optimization stages are highly consistent, reflecting the intrinsic homogeneity between constructing semantic representations and training the clustering network — a perspective previously overlooked in TIC research. By transitioning from empirical heuristics to a unified, principled optimization paradigm, our algorithm advances the development of TIC toward a more theoretically grounded direction.

Algorithm 1 Our Algorithm (S^3)

Input: Image encoder f, text encoder g with parameters θ; all nouns $\mathcal{T}^{\text{all}}$ from WordNet; unlabeled dataset $\mathcal{D}$.
Output: Optimized nouns $\hat{\mathcal{T}}$ and model parameters $\hat{\theta}$.

Randomly initialize $\mathcal{T} = \{t_1, \ldots, t_K\} \subset \mathcal{T}_{\text{all}}$.
Load pre-trained CLIP model parameters θ.
// Stage 1: Semantic Representations Construction
// —— Coordinate Descent ——
Pre-extract text embeddings of $\mathcal{T}^{\text{all}}$ and image embeddings of $\mathcal{D}$.
Pre-perform nearest-neighbor search based on cosine similarity.
for $s = 1$ to N_{iter} **do**
  for $j = 1$ to K **do**
    Randomly sample mini-batch $\mathcal{B}_j^{(s)} \subset \mathcal{D}$ with N_{image} samples.
    Randomly sample mini-batch $\mathcal{T}_j^{(s)} \subset \mathcal{T}^{\text{all}}$ with N_{text} nouns.
    for each $t_j \in \mathcal{T}_j^{(s)} \cup \{t_j^{(s)}\}$ **do**
      Construct $\mathcal{T}'$ using t_j according to Eq. (8).
      Compute $\mathcal{L}_{\text{principle}}(\mathcal{T}', \mathcal{B}_j^{(s)})$ according to Eq. (12).
    end for
    $\hat{t}_j \leftarrow \arg\min \mathcal{L}_{\text{principle}}$.
    $t_j^{(s+1)} \leftarrow \hat{t}_j$.
  end for
end for
Obtain the optimal noun set $\hat{\mathcal{T}}$.
// Stage 2: Clustering Network Training
// —— Gradient Descent ——
for iteration **do**
  Sample mini-batch $\mathcal{B} \subset \mathcal{D}$.
  Obtain two augmented views for all images in $\mathcal{B}$.
  Extract image embeddings via f.
  Extract text embeddings for all nouns in $\hat{\mathcal{T}}$ via g.
  Compute losses according to Eq. (12) and Eq. (15).
  Update θ using gradient descent.
end for
Obtain the optimal $\hat{\theta}$.
return $\hat{\mathcal{T}}, \hat{\theta}$.

B Convergence Analysis of Coordinate Descent

We provide a formal convergence analysis for the coordinate descent stage in Eq. (8). The analysis is intentionally stated for the semantic representation construction stage only: the second stage is a standard gradient-based network training procedure, while the first stage is the discrete optimization component that may raise convergence concerns.

Let $\Omega = \{\mathcal{T} = (t_1, \ldots, t_K) : t_j \in \mathcal{T}^{all}, t_i \neq t_j \text{ for } i \neq j\}$ denote the feasible set of non-overlapping semantic representations. Since $\mathcal{T}^{all}$ is finite, Ω is also finite. For notational simplicity, define the full-data objective

$$F(\mathcal{T}) = \mathcal{L}_{\text{principle}}(\mathcal{T}, \mathcal{D}). \tag{16}$$

Given $\mathcal{T} = (t_1, \ldots, t_K)$, let $\mathcal{T}_{j \leftarrow u}$ denote the set obtained by replacing the j-th coordinate with u while keeping all other coordinates fixed. We call $\mathcal{T}^\star$ a coordinate-wise local minimum if

$$F(\mathcal{T}^\star) \leq F(\mathcal{T}^\star_{j \leftarrow u}), \quad \forall j \in \{1, \ldots, K\},\ \forall u \in \mathcal{T}^{all}. \tag{17}$$

Lemma 1 (Bounded finite objective)**.** *The objective $F(\mathcal{T})$ is lower bounded on Ω. Moreover, since Ω is finite, F is Lipschitz continuous under the discrete metric on Ω.*

Proof. For any image, the prediction vector $\mathbf{p}_i$ lies in the probability simplex. Therefore, $0 \leq \mathcal{L}_{ent}(\mathbf{x}_i) \leq \log K$. For the neighborhood term, each inner product between two probability vectors lies in $[0, 1]$, and the number of nearest neighbors is fixed; hence $\mathcal{L}_{nn}$ is bounded from below by a finite constant. For the diversity term, $\bar{\mathbf{p}}$ is also in the probability simplex, which gives $-\log K \leq \mathcal{L}_{div} \leq 0$. Consequently, $\mathcal{L}_{S^3} = \mathcal{L}_{ent} + \mathcal{L}_{nn} + \lambda_{div}\mathcal{L}_{div}$ is lower bounded. Since Eq. (13) sets $\mathcal{L}_{\text{principle}} = \mathcal{L}_{S^3}$ in our method, F is lower bounded. Finally, any function over a finite metric space is Lipschitz continuous under the discrete metric with a finite Lipschitz constant. □

Proposition 2 (Finite convergence of full-batch coordinate descent)**.** *Consider the deterministic version of Eq.* (8)*, where $\mathcal{B}_j^{(s)} = \mathcal{D}$ and the full vocabulary $\mathcal{T}^{all}$ is searched at each coordinate update. Assume ties are broken in favor of the current coordinate value $t_j^{(s)}$. Then cyclic coordinate descent monotonically decreases F and reaches a coordinate-wise local minimum after finitely many strict updates.*

Proof. At a coordinate update, the current solution $\mathcal{T}^{(s)}$ is included in the search set because $t_j^{(s)}$ is a feasible candidate. The update therefore satisfies

$$F(\mathcal{T}^{(s+1)}) \leq F(\mathcal{T}^{(s)}). \tag{18}$$

With the stated tie-breaking rule, the iterate changes only when a strict decrease is achieved. Since Ω is finite, a sequence of strict decreases cannot visit infinitely many distinct states. Thus, after finitely many strict updates, no coordinate replacement can further decrease F. By definition, the resulting solution is a coordinate-wise local minimum. □

We next discuss the stochastic implementation used in Algorithm 1, where both images and noun candidates are sampled for efficiency. Let $\widehat{F}_{\mathcal{B}}(\mathcal{T}) = \mathcal{L}_{\text{principle}}(\mathcal{T}, \mathcal{B})$ denote the minibatch empirical objective. The following assumption is standard for finite-sample empirical objectives over a finite candidate class.

Assumption 3 (Uniform minibatch accuracy)**.** *For any sampled coordinate candidate set $\mathcal{C} \subset \Omega$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\sup_{\mathcal{T} \in \mathcal{C}} \left| \widehat{F}_{\mathcal{B}}(\mathcal{T}) - F(\mathcal{T}) \right| \leq \varepsilon(N_{image}, |\mathcal{C}|, \delta), \tag{19}$$

where $\varepsilon(N_{image}, |\mathcal{C}|, \delta) = O\left(\sqrt{\log(|\mathcal{C}|/\delta)/N_{image}}\right)$.

Remark 4. Assumption 3 follows from boundedness and concentration over a finite class. The boundedness is guaranteed by Lemma 1; the finite class is the sampled coordinate candidate set. Batch-level terms such as $\mathcal{L}_{div}$ may introduce a negligible finite-sample bias, which is absorbed into ε.

Proposition 5 (Approximate descent under minibatch sampling)**.** *Let $\mathcal{C}_j^{(s)}$ be the sampled candidate set for coordinate j at step s, and let $\mathcal{T}^{(s+1)}$ be selected by minimizing $\widehat{F}_{\mathcal{B}_j^{(s)}}$ over $\mathcal{C}_j^{(s)}$. On the event in Assumption 3, for every $\mathcal{U} \in \mathcal{C}_j^{(s)}$,*

$$F(\mathcal{T}^{(s+1)}) \le F(\mathcal{U}) + 2\varepsilon(N_{image}, |\mathcal{C}_j^{(s)}|, \delta). \tag{20}$$

In particular, since the current solution is included in $\mathcal{C}_j^{(s)}$,

$$F(\mathcal{T}^{(s+1)}) \le F(\mathcal{T}^{(s)}) + 2\varepsilon(N_{image}, |\mathcal{C}_j^{(s)}|, \delta). \tag{21}$$

Proof. By the definition of the update,

$$\widehat{F}_{\mathcal{B}_j^{(s)}}(\mathcal{T}^{(s+1)}) \le \widehat{F}_{\mathcal{B}_j^{(s)}}(\mathcal{U}), \quad \forall \mathcal{U} \in \mathcal{C}_j^{(s)}. \tag{22}$$

Applying Assumption 3 to both $\mathcal{T}^{(s+1)}$ and $\mathcal{U}$ gives

$$F(\mathcal{T}^{(s+1)}) \le \widehat{F}_{\mathcal{B}_j^{(s)}}(\mathcal{T}^{(s+1)}) + \varepsilon \le \widehat{F}_{\mathcal{B}_j^{(s)}}(\mathcal{U}) + \varepsilon \le F(\mathcal{U}) + 2\varepsilon, \tag{23}$$

where $\varepsilon = \varepsilon(N_{\text{image}}, |\mathcal{C}_j^{(s)}|, \delta)$. Taking $\mathcal{U} = \mathcal{T}^{(s)}$ yields the second inequality. □

Corollary 6 (Neighborhood of a coordinate-wise local minimum)**.** *Under Assumption 3, the stochastic coordinate update detects every sampled coordinate improvement whose full-objective margin is larger than 2ε. Therefore, the stochastic procedure converges to an $O(\varepsilon)$ neighborhood of a coordinate-wise local minimum, up to the probability of not sampling improving noun candidates.*

Proof. Suppose there exists a sampled candidate $\mathcal{U} \in \mathcal{C}_j^{(s)}$ such that

$$F(\mathcal{U}) \le F(\mathcal{T}^{(s)}) - \eta \quad \text{for some } \eta > 2\varepsilon. \tag{24}$$

By Proposition 5,

$$F(\mathcal{T}^{(s+1)}) \le F(\mathcal{T}^{(s)}) - (\eta - 2\varepsilon), \tag{25}$$

so any sampled improvement larger than the minibatch uncertainty yields a strict full-objective decrease. If no such improvement exists, then the current solution is already $O(\varepsilon)$-coordinate-wise optimal over the sampled candidate set. For the full vocabulary, the only additional failure mode is that the improving noun is not sampled. If m nouns yield an η-improvement for a coordinate and $M = |\mathcal{T}^{all}|$, sampling N_{text} nouns includes at least one of them with probability at least

$$1 - \left(1 - \frac{m}{M}\right)^{N_{\text{text}}}, \tag{26}$$

which increases monotonically with N_{text}. Repeating cyclic updates for N_{iter} iterations further reduces the probability of consistently missing improving candidates. Thus, larger N_{image} shrinks the optimization neighborhood through a smaller ε, while larger N_{text} and N_{iter} increase the probability of reaching that neighborhood. □

This analysis complements the empirical convergence evidence in Fig. 2 and the hyperparameter sensitivity study in Fig. 3.

C Detailed Discussion of Existing Semantic Construction Strategies

Existing TIC methods generally follow a two-stage paradigm: they first construct semantic representations from a large candidate text space, and then use these representations to enrich discriminative signals for clustering network training. Despite their empirical effectiveness, the first-stage semantic construction procedures in TAC (Li et al., 2024), SIC (Cai et al., 2023), and MCA (Qiu et al., 2024) are largely heuristic. In this section, we provide a more detailed discussion of their limitations and explain why a global optimization perspective is necessary.

Element-wise handcrafted priors. Some existing methods introduce priors such as uniqueness (Cai et al., 2023) and hierarchy (Qiu et al., 2024) to filter candidate nouns from $\mathcal{T}^{all}$. These strategies can be viewed as optimizing or filtering each candidate independently, i.e., through a criterion of the form $\mathcal{L}_{\text{principle}}(t_i)$. Although such priors are useful for removing easy negative cases, they do not explicitly evaluate whether a set of selected semantic representations $\mathcal{T}$ collectively explains the structure of the image dataset $\mathcal{D}$. Consequently, they may retain candidate texts that appear plausible in isolation but are not useful for separating the actual clusters. Moreover, these handcrafted rules often introduce several data-specific hyperparameters, such as filtering thresholds, hierarchy constraints, or candidate-selection ratios, which makes the first stage less principled and independent of the second-stage clustering objective.

Instance-level point-to-point matching. Another common strategy is instance-level point-to-point (P2P) cross-modal matching, where each image embedding is directly matched to nearby text embeddings in the VLM embedding space. This strategy is vulnerable to imperfect image-text alignment (Radford et al., 2021; Singh & Zheng, 2023; Zhang et al., 2024; Abbasi et al., 2025). The core goal of TIC is to identify text-modal information relevant to the dataset in order to enhance image clustering. However, the candidate text space contains a large amount of irrelevant text, and the image-text alignment of CLIP is not perfect. Existing point-wise strategies struggle to handle such noise because isolated points cannot precisely characterize semantic boundaries. For example, consider an image of "a dog on the grass." It may have high similarity with both "grass" and "dog." Alternatively, due to imperfect alignment, it might even be considered similar to "cat." When matching is performed only at the level of isolated image-text pairs, such ambiguous or incorrect local similarities can be directly injected into the semantic representations, introducing noise detrimental to clustering.

Center-based category mapping. To obtain category-level semantic representations, TAC (Li et al., 2024) and SIC (Cai et al., 2023) use K-means centroids or similar visual centers as intermediate anchors for text matching. This reduces the influence of individual noisy samples, but it also represents a complex image embedding manifold using a small set of centers. As a result, substantial intra-class structural information is discarded. For instance, dog images may appear under different backgrounds, poses, and breeds; representing them by a single center loses the similarity relationships among these diverse subgroups. More generally, center-based mapping models complex class distributions using centroids, thereby weakening the global intra-class and inter-class structural information that is crucial for reliable semantic boundary identification.

Why global structure matters. The limitation of the above strategies lies in their local perspective. They either judge each candidate text independently, match isolated image-text pairs, or compress the image distribution into coarse centers. However, semantic relevance is often only identifiable from the global feature structure of the whole dataset. Returning to the example above, an individual dog image may be locally similar to "grass," "dog," or even "cat." If we consider the entire dataset, however, we may observe many dog images across different environments that are mutually similar, as well as true cat images that are more similar to the text "cat" than any dog image. This global evidence provides a more reliable basis for identifying candidate texts and distinguishing semantic boundaries.

Therefore, we argue that existing first-stage strategies are constrained by two forms of locality. First, they model complex class distributions using isolated candidates, image-text pairs, or visual centers, thereby losing global intra-class structural information. Second, they perform semantic matching without fully considering global intra-class and inter-class relationships. Given the strong correlation between semantic representation quality and final clustering performance (Qiu et al., 2024; Cai et al., 2023; Li et al., 2024), these limitations can lead to suboptimal clustering results. In contrast, S^3

directly formulates semantic representation construction as a dataset-level optimization problem, where the selected semantic representations are evaluated by how well they align with the global feature structure of $\mathcal{D}$. This shifts the first stage of TIC from handcrafted local matching toward principled global optimization.

D STABILITY OF RESULTS

Table 7 reports the sample variance of our algorithm over three runs.

Table 7: Performance metrics (ACC, NMI, ARI) with standard deviations on various datasets.

	STL10			**CIFAR10**			**CIFAR20**			**ImageNet-Dogs**			**Tiny-ImageNet**			**DTD**			**UCF101**		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
std.	0.13	0.31	0.27	0.08	0.17	0.17	1.68	1.43	1.72	1.51	0.99	2.07	1.75	0.69	1.47	2.01	1.53	2.13	1.40	0.69	1.44

E COMPUTATIONAL EFFICIENCY.

Our algorithm incorporates a coordinate descent optimization stage, which, to our knowledge, has not been previously applied in the TIC methods. To address potential concerns regarding the computational efficiency of the coordinate descent stage, we measured the runtime of both stages. On the DTD dataset, which contains 47 categories, our first stage (coordinate descent) executes in approximately 125 seconds. In comparison, the second stage (clustering network training) requires 1,276 seconds, significantly longer than the first stage. The experiments were conducted on a single RTX A6000 GPU.

F VALIDATING THE ROLE OF FEATURE STRUCTURE

In this experiment, we validate the importance of feature structure for obtaining high-quality semantic representations. To this end, we design a "variant" algorithm in which the full dataset $\mathcal{D}$ in Eq. (6) is replaced with the K-means cluster centroids $\mathbf{u}$, while all other procedures remain unchanged. Notably, the core characteristic of this "variant" is that it uses the cluster centroids to represent the global feature structure, which is similar to the approach in the existing TIC methods. Table 8 presents the results, showing that our standard method significantly outperforms this "variant", thereby confirming our hypothesis.

Table 8: Impact of feature-structure guidance on semantic representations. Comparison between our full S^3 method and a variant using K-means centroids in Eq. (6), evaluated on CIFAR20, ImageNet-Dogs, and Tiny-ImageNet (ACC, NMI, ARI).

Method	**CIFAR20**			**ImageNet-Dogs**			**Tiny-ImageNet**		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Variant	53.6	58.3	39.6	69.3	72.8	60.3	54.3	76.0	44.8
S^3 (Ours)	**62.8**	**65.6**	**49.6**	**86.0**	**86.5**	**79.8**	**67.9**	**82.6**	**57.7**

G LIMITATIONS AND FUTURE WORK.

While we have established a principled optimization framework for TIC, our current implementation prioritizes simplicity and reproducibility by employing basic strategies such as random noun sampling and cyclic coordinate descent. We acknowledge that these elementary choices may potentially yield suboptimal solutions. Consequently, we envision several promising avenues for future improvement. For instance, optimization efficiency could be enhanced by designing a scoring function to preferentially sample text candidates most pertinent to the current dataset, or by developing an adaptive strategy to identify and update the specific semantic representation t_j that offers the greatest gain at each step. Furthermore, the intrinsic homogeneity between the two-stage optimization objectives—preliminarily verified in this work—warrants more in-depth theoretical exploration, which could potentially pave the way for a fully unified TIC paradigm.

H CONCRETE EXTENSION DIRECTIONS

Efficient candidate sampling. As shown in Fig. 3, the current random sampling strategy with $N_{\text{text}} = 5000$ is already close to performance saturation. This is intuitive, since with over 60k nouns in total, sampling 5,000 each time is unlikely to miss relevant concepts. However, this introduces substantial computational cost. One concrete improvement is to use content-aware sampling: when $N_{iter} \geq 1$, suppose we are optimizing t_j, and let $\mathbf{c}_j$ denote the center of the samples predicted to belong to t_j in the previous iteration. We then sample the 250 nouns in $\mathcal{T}^{all}$ closest to $\mathbf{c}_j$. Experimental results show that this strategy has negligible impact on final performance but reduces the Stage 1 runtime on Tiny-ImageNet by 51%.

I DISCUSSION ON CORE INNOVATIONS

The primary innovation of this work lies in advancing Text-assisted Image Clustering (TIC) from empirical heuristics toward a principled optimization framework.

First, we provide a unified formalization of the TIC paradigm, modeling it as a two-stage optimization problem. Crucially, we propose solving the semantic representation construction from a global perspective, rather than relying on local, instance-level matching. This global view is essential for mitigating the noise inherent in imperfect vision-language alignment, and we effectively tackle the resulting combinatorial complexity via a coordinate descent algorithm. Both this **rigorous formalization** and the **global optimization perspective** constitute contributions to the TIC.

Second, within this framework, we achieve semantic representation construction using an objective function that is **homogeneous** to the subsequent clustering network training. This revealed homogeneity prompts a profound theoretical question: *Could semantic construction and clustering network training fundamentally be manifestations of the same underlying problem?* This suggests that advanced TIC strategies could potentially be derived directly from established, high-performing IC methodologies. While we do not claim to fully resolve this unification herein, our work provides a foundational framework that opens this avenue for future research.

Finally, it is vital to clarify that our contribution extends beyond a mere combination of existing loss functions. We innovate by repurposing objectives traditionally designed for *optimizing clustering assignments* to instead *select semantic representations*. This methodological leap is not arbitrary but is intuitively derived from our proposed principle that Structure Shapes Semantics (S^3), demonstrating how the intrinsic feature structure of a dataset can be leveraged to rigorously constrain and optimize semantic representations.

J Robustness to Misleading Concepts

In this section, we conduct quantitative experiments to evaluate the robustness of our algorithm when the candidate vocabulary pool $\mathcal{T}^{all}$ contains a substantial number of misleading concepts. To systematically construct a candidate set populated with such specific distractors, we crawled all 557 dog breeds listed on the *Wikipedia* page "List of dog breeds". We utilize the ImageNet-Dogs dataset for this evaluation. In this setting, the crawled breeds are highly relevant to the dataset domain but introduce significant ambiguity due to the fine-grained nature of dog breeds. Consequently, these concepts are visually and semantically prone to confusion, serving as rigorous misleading concepts to test the discriminative capability of our optimization framework.

Table 9: Robustness comparison under misleading concept vocabulary on ImageNet-Dogs. Evaluation is performed prior to clustering network training.

Vocabulary	ACC	NMI	ARI
WordNet (Ours)	72.7	73.4	59.9
Misleading Concepts	**80.1**	**79.3**	**68.6**
GT Class Names	72.8	73.5	58.2

As shown in Table 9, remarkably, the performance using Misleading Concepts significantly surpasses that of using both WordNet and GT class names. This observation may appear counter-intuitive at first glance. However, it is crucial to recognize that the core objective of Text-assisted Image Clustering is to effectively distinguish between different image clusters, rather than to assign precise fine-grained textual labels to them. Unlike WordNet, which contains a vast amount of semantically irrelevant noisy nouns, the Misleading Concepts set confines the candidate pool to a domain-specific scope—specifically, dog breeds—making nearly all candidates highly relevant to the ImageNet-Dogs dataset. Furthermore, in terms of capturing inter-class differences, the Misleading Concepts provide a finer-grained representation of dog varieties compared to WordNet, thereby facilitating better separation of distinct categories. In summary, although establishing precise correspondences between clusters and specific semantic labels is challenging due to the presence of misleading concepts, the dense and fine-grained nature of this vocabulary proves advantageous for the primary task of image clustering—separating visual groups.

K ROBUSTNESS TO NOISY CONCEPTS

In this subsection, we quantitatively analyze the robustness of our algorithm when the candidate vocabulary $\mathcal{T}^{all}$ is heavily contaminated with dataset-irrelevant noisy concepts. To construct a substantial pool of such distractors, we crawled all 557 entries from the Wikipedia "List of dog breeds" page and replicated them 20 times, resulting in a total of 11,140 items. We conduct experiments on CIFAR10 and Tiny-ImageNet, two datasets comprising general object categories. In this context, the fine-grained dog breeds are semantically orthogonal to the target concepts, thereby constituting a semantic space with significant noise. To demonstrate the efficacy of our method under such adverse conditions, we construct the candidate set by mixing the ground-truth category names $\mathcal{T}^{GT}$ with the noise set, formulated as $\mathcal{T}^{all} = \mathcal{T}^{GT} \cup \mathcal{T}^{noise}$. A robust algorithm is expected to successfully recover $\mathcal{T}^{GT}$ from the mixture while filtering out the irrelevant noise. To simulate varying noise intensities, we further sample $N_{noise} \in \{0, 100, 1000, 10000\}$ concepts from the noise pool.

Table 10: Robustness to noisy vocabularies on CIFAR10. Evaluation is performed prior to clustering network training.

Lexicon	**CIFAR10**		
	ACC	NMI	ARI
$\lvert\mathcal{T}^{GT}\rvert$=10, $\lvert\mathcal{T}^{noise}\rvert$=0	90.0	80.7	79.3
$\lvert\mathcal{T}^{GT}\rvert$=10, $\lvert\mathcal{T}^{noise}\rvert$=100	89.8	80.3	79.0
$\lvert\mathcal{T}^{GT}\rvert$=10, $\lvert\mathcal{T}^{noise}\rvert$=1000	89.7	80.4	78.9
$\lvert\mathcal{T}^{GT}\rvert$=10, $\lvert\mathcal{T}^{noise}\rvert$=10000	89.7	80.3	78.7

Table 11: Robustness to noisy vocabularies on Tiny-ImageNet. Evaluation is performed prior to clustering network training.

Lexicon	**Tiny-ImageNet**		
	ACC	NMI	ARI
$\lvert\mathcal{T}^{GT}\rvert$=200, $\lvert\mathcal{T}^{noise}\rvert$=0	81.7	86.4	68.1
$\lvert\mathcal{T}^{GT}\rvert$=200, $\lvert\mathcal{T}^{noise}\rvert$=100	78.6	85.8	66.5
$\lvert\mathcal{T}^{GT}\rvert$=200, $\lvert\mathcal{T}^{noise}\rvert$=1000	78.0	85.8	66.1
$\lvert\mathcal{T}^{GT}\rvert$=200, $\lvert\mathcal{T}^{noise}\rvert$=10000	77.2	85.5	65.4

As evidenced in Tables 10 and 11, our method maintains remarkably stable performance on both datasets even as $|\mathcal{T}^{noise}|$ increases orders of magnitude. Notably, on CIFAR10 with $|\mathcal{T}^{noise}| = 10,000$—where the signal-to-noise ratio drops to a mere 0.001—the accuracy declines by only $\approx$0.3%, confirming that our global structure-guided optimization is highly resilient to semantic noise.

L BOOSTING PERFORMANCE VIA BETTER INITIALIZATION

In this section, we scrutinize the robustness of our optimization framework with respect to the initialization of the semantic candidate set $\mathcal{T}$. While our default approach employs random initialization to demonstrate the efficacy of the coordinate descent algorithm in navigating the solution space without priors, we acknowledge that providing a "warm start" could potentially accelerate convergence or lead to better local optima. To systematically evaluate this, we compare three distinct initialization strategies:

1. **Random Initialization**, which serves as the baseline to test the intrinsic optimization capability of our S^3 principle;
2. **Centroid-based Linear Assignment**, a heuristic strategy where we first perform K-means (MacQueen et al., 1967) on image embeddings to obtain visual centroids and then apply the Hungarian algorithm to compute an optimal linear assignment between these centroids and the candidate textual semantic space $\mathcal{T}^{all}$;
3. **Ground-Truth Initialization**, where $\mathcal{T}$ is initialized using the ground-truth class names to establish the theoretical performance upper bound achievable through optimal initialization.

By assessing the performance gaps across these settings, we aim to verify whether our structure-guided optimization is robust enough to rectify suboptimal starting points and to quantify the potential ceiling of semantic representation quality.

Table 12: Comparison of clustering performance using different initialization strategies. **Random**: Random initialization (Ours). **Linear Assignment**: Centroid-based image-text linear assignment. **Ground Truth**: Upper bound using GT class names.

Initialization	CIFAR20			ImageNet-Dogs		
	ACC	NMI	ARI	ACC	NMI	ARI
Random (Ours)	62.8	65.6	49.6	86.0	86.5	79.8
Linear Assignment	64.0	66.1	50.6	86.3	86.7	79.7
Ground Truth	**65.0**	**67.5**	**51.9**	**86.9**	**87.1**	**80.0**

The quantitative results are summarized in Table 12. Remarkably, our method using Random initialization already achieves impressive performance, reaching 62.8% and 86.0% ACC on CIFAR20 and ImageNet-Dogs, respectively. Critically, the performance gap between this random baseline and the Ground Truth upper bound is marginal (e.g., merely 0.9% difference on ImageNet-Dogs). This empirical evidence strongly validates the robustness of our optimization framework, demonstrating that the S^3 principle can effectively navigate the solution space to locate high-quality semantic representations without relying on warm starts. While the Linear Assignment strategy offers a slight performance gain, moving closer to the theoretical limit, the substantial efficacy of the random initialization confirms that our framework's power stems primarily from the principled optimization itself. Nevertheless, bridging the remaining gap through advanced initialization strategies remains a potential avenue for future improvement.

M BOOSTING PERFORMANCE VIA BETTER LEXICON

To further analyze the potential and upper bound of our method, we replace WordNet (Miller, 1995) with a curated set of 22k visually relevant nouns (denoted as Visual Nouns (Zheng et al., 2024)) and the ground truth class names (Ground Truth) for experiments with S^3. When using Ground Truth, our algorithm bypasses Eq. (6) and directly executes Eq. (7). As shown in Table 13, the vocabulary gradually decreases in size and becomes increasingly aligned with the datasets from top to bottom. Correspondingly, performance improves with the quality of the vocabulary. Notably, on ImageNet-dogs, performance nearly reaches the GT level, demonstrating the effectiveness of our optimization.

Table 13: Comparison of clustering performance using different vocabularies.

Lexicon	**CIFAR20**			**ImageNet-Dogs**		
	ACC	NMI	ARI	ACC	NMI	ARI
WordNet (Miller, 1995)	62.8	65.6	49.6	86.0	86.5	79.8
Visual Nouns (Zheng et al., 2024)	65.5	67.5	51.9	87.3	**86.8**	**80.9**
Ground Truth	**72.2**	**70.5**	**56.3**	**88.7**	85.6	80.2

N Additional Generalization Studies

We further evaluate S^3 in more challenging scenarios, including cross-domain transfer, long-tail distributions, and fine-grained classification. Since existing TIC methods have not been evaluated on these datasets, the reported results are based on our own reproductions. Following the hyperparameter settings used in TAC for other datasets, we set the number of K-means clusters to three times the number of ground-truth classes. For S^3 (Ours), we fix τ_{co} to the same value as on UCF101 without tuning any other hyperparameters.

Table 14: Additional ACC results on cross-domain, long-tail, and fine-grained datasets.

Method	OfficeHome Avg	CIFAR100-LT-IF10	CIFAR100-LT-IF50	CIFAR100-LT-IF100	CUB
TAC	68.07	49.20	39.92	38.05	45.03
S^3 (Ours)	**73.90**	**57.25**	**49.42**	**48.14**	**53.19**

As shown in Table 14, our method significantly outperforms TAC in cross-domain, long-tail, and fine-grained classification. This result is also intuitive, since our improvement mainly comes from enhancing the quality of the text representations obtained in the first stage. Higher-quality text representations generally lead to better clustering performance in most scenarios.

Specially, the CIFAR100-LT results provide direct evidence under class imbalance: although $\mathcal{L}_{div}$ introduces a balanced-clustering preference, it does not prevent S^3 from obtaining clear gains on long-tail data. In fact, $\mathcal{L}_{div}$ has also been widely adopted in existing TIC methods (Li et al., 2024; Cai et al., 2023; Qiu et al., 2024). It imposes a balancing regularization over clustering assignments to ensure that training does not collapse, since "assigning all images to a single cluster" is a degenerate solution for the other objectives. Even under long-tail settings, this balancing regularization remains important for maintaining training stability.

O CROSS-MODEL CONSISTENCY

The main experiments instantiate our framework with CLIP, following the standard TIC setting. To examine whether the S^3 principle is tied to a specific VLM, we additionally evaluate the method using BLIP (Li et al., 2022a).

Table 15: ACC results using BLIP.

Method	CIFAR20	DTD
TAC	61.53	54.37
S^3 (Ours)	**64.11**	**57.82**

As shown in Table 15, S^3 also performs well on BLIP (Li et al., 2022a). This is consistent with our intuition that the principle itself is sufficiently fundamental: feature structure is crucial for shaping textual semantic representations, regardless of the specific VLM used to provide the initial image-text embedding space. These results provide preliminary evidence of cross-model consistency.

P DISCUSSION ON UCF101

Our method achieves comprehensive advantages on six out of seven datasets in the main results; on UCF101, it achieves the best NMI but is slightly behind TAC in ACC and ARI. This gap may be due to the use of class-level semantic representations in methods such as S^3 and SIC, rather than instance-level representations as in TAC. Action recognition in UCF101 contains large intra-class appearance variation and context-dependent visual cues, where instance-level textual supervision may provide a small advantage.

Nevertheless, the UCF101 result should be interpreted together with the broader evidence. Compared with SIC, although we also use class-level semantic representations, the quality of our semantic representations is higher due to global optimization and feature-structure guidance. Therefore, S^3 achieves ACC improvements of 4.7% and 6.4% over SIC on DTD and UCF101, respectively. Compared with TAC, although TAC slightly surpasses us on UCF101 with a 0.4% ACC advantage, it performs worse than our method on the other six datasets. This suggests that global optimization is broadly beneficial for category-level semantic construction, while instance-level semantic representations remain a promising direction for datasets with highly diverse actions or contexts.

Q VISUALIZATIONS OF WORDNET NOUNS

To illustrate the complexity of the candidate pool from which our framework operates, we visualize random samples of nouns from the WordNet lexicon. We conducted three independent trials, randomly selecting 30 nouns in each instance. As shown in the following lists, these nouns are largely irrelevant to any specific image dataset due to the nature of random sampling. By presenting these samples, we emphasize the significant semantic noise present in the massive candidate pool. This visualization serves to demonstrate the robustness of our principled optimization, which successfully extracts high-quality, dataset-relevant semantic representations even when starting from such a noisy and diverse source.

Crucially, the utilization of WordNet ensures that our approach is not constrained by dataset-specific priors. As a general-purpose lexicon, it introduces no domain bias, thereby avoiding the limitations often associated with restricted or handcrafted vocabularies. This universality is quantitatively validated by our extensive experiments across seven diverse benchmarks. Our method demonstrates exceptional robustness not only on specialized datasets such as DTD (textures) and UCF-101 (actions) but also on challenging, large-scale general datasets like Tiny-ImageNet. These results confirm that our principled optimization framework can effectively distill high-quality, dataset-relevant semantics from a vast, noise-heavy generic lexicon without relying on prior knowledge.

Randomly selected nouns from WordNet

looking-glass plant, penelope, muff, lady tulip, cuddy, endogeny, thuggery, genus origanum, silverware, pancreas, suffragism, abolition, patras, ground loop, bunyan, spring squill, harpsichordist, 'hood, tampere, selfsameness, notoryctus, oncidium, policeman bird, foreign agent, nurse-midwife, temporal bone, orthilia, travelogue, cranesbill, earldom.

Randomly selected nouns from WordNet

heat of solution, bristle fern, bitterroot, cos, carbon black, cubby, wild china tree, probenecid, bounder, purple silkweed, aix, anaximander, gilbert islands, ascoma, carina, police commissioner, price increase, collectivist, wister, chain letter, continuity irish republican army, heliopause, pasch, meadow salsify, adjoining room, weissbier, diabetes insipidus, air transportation system, pulmonary artery, hypersensitivity reaction.

Randomly selected nouns from WordNet

seder, ox, campyloneurum, switch, pinball machine, paraguay, turtle, succotash, nitrous oxide, rice weevil, brickyard, palestine liberation organization, shower curtain, lesser wintergreen, axial skeleton, birchbark canoe, gyrfalcon, strip mine, direct support, wild potato, paresthesia, tipulidae, sulfapyridine, hole-in-the-wall, motif, caftan, pakistani, bagassosis, euphausia pacifica, seychellois.

R VISUALIZATIONS OF OPTIMIZED NOUNS

This section provides a qualitative visualization of the category-level textual semantic representations $\hat{\mathcal{T}}$ obtained through the proposed optimization framework. We present the specific nouns selected by our method for all datasets to illustrate how the global feature structure identifies semantically coherent and visually grounded concepts. These examples demonstrate the framework's ability to recover target categories with high fidelity, even when facing the challenges of imperfect cross-modal alignment.

Optimized nouns for STL10

squirrel monkey, american saddle horse, aircraft landing, cuculiform bird, school ship, seizure-alert dog, venison, used-car, european wildcat, garbage truck.

Optimized nouns for CIFAR10

percheron, sedan, trailer truck, egyptian cat, japanese spaniel, caprimulgiform bird, waterbuck, tarahumara frog, guard boat, reconnaissance plane.

Optimized nouns for CIFAR20

eastern lowland gorilla, pacific bottlenose dolphin, tree farm, artificial flower, lion cub, jerboa rat, bedroom furniture, long-horned beetle, giant eland, motor vehicle, shot glass, arabian sea, cn tower, simple fruit, au pair girl, geastrum coronatum, clock radio, ringneck snake, tudor architecture, soldierfish.

Optimized nouns for DTD

cherimoya, single crochet, amethystine python, orb web, checkerboard, paisley, perfoliate leaf, pothole, glove leather, quartz crystal, hubbard squash, waffler, acne rosacea, gneiss, cork tree, desert soil, chicken wire, artificial flower, stinging hair, finger-painting, sugariness, candy striper, zebra, spotted weakfish, harris tweed, side dish, spirula, square knot, yellow-green algae, heat exchanger, bread dough, lascaux, barnyard grass, bamboo curtain, balldress, notepad, foil, garden lettuce, lecanora, sea spray, white leather, millboard, chevron, sparkling wine, polka dot, dry point, apiary.

Optimized nouns for UCF101

chocolate sauce, track event, cricket match, cymbal, duckpin, widegrip pushup, plagiocephaly, ice skate, hindustani, professional boxing, foul shot, bandoneon, inning, professional tennis, marching band, corner kick, sitar player, rock climbing, darning needle, cow pony, tap dancer, round of golf, bat mitzvah, guitarist, bicycling, military ceremony, kayak, piano sonata, vector algebra, skydiving, after-shave, menuhin, touch typing, sumo ring, sawhorse, trampoline, hairdressing, birthday cake, crossbow, billiards, purple onion, skateboarding, table-tennis table, incline bench press, skiing, floccose chanterelle, white water, rowing, headstand, volleyball player, table knife, touch football, rastafari, toll call, social dancing, electric toothbrush, housework, field hockey, leash, prostration, clarinetist, trapeze, gardener, surfing, archery, drummer, aerobics, ladder truck, sled dog, backstroke, gymnastics, discus, pillow fight, bottomland, hula, makeup, self-rising flour, punching bag, cello, epee, playground, soffit, cliff diving, batting cage, thoroughbred racing, food bank, barber, blackboard, smoke bomb, elevator man, english setter, trailer camp, table rapping, basketball court, water-skiing, dressing room, kung fu, curling iron, clean and jerk, preschooler, toddler.

Optimized nouns for Tiny-ImageNet

alouatta, shad roe, umbrella, sanitary landfill, hurricane lamp, fly rod, tamale pie, cantilever bridge, refrigerator, limousine, queensboro bridge, diesel locomotive, freight elevator, rood screen, sports car, tractor, meadow mushroom, analog clock, stockinette stitch, valencia orange, sicilian pizza, black-tailed deer, persian cat, over-the-counter market, norway lobster, hourglass, christmas stocking, old english sheepdog, pay-phone, ao dai, robusta coffee, arabian camel, volleyball net, genus jabiru, triumphal arch, morris dancer, cliff dwelling, flat-coated retriever, terraced house, parking meter, sombrero, graffito, viaduct, delivery truck, shift key, go-kart, dugong, tape player, punching bag, lifeboat, pikeblenny, spillway, staghorn coral, lesser panda, swimming trunks, lioness, mortarboard, water tower, uncompahgre peak, bullet train, goldfish, genus capsicum, bathtub, comic book, abacus, northern cricket frog, chalcopyrite, video ipod, scoreboard, cash machine, clarinet, speckled rattlesnake, flag waving, bofors gun, european fire salamander, house centipede, grand canal, rocking chair, electric frying pan, malinois, ichneumon fly, lighthouse keeper, goosander, leopard cat, pomegranate, pill bottle, african elephant, banana quit, slide rule, pedicab, school bus, golden retriever, beach erosion, air-to-air missile, rugby, sunglass, grape arbor, ice-cream sundae, common duckweed, king penguin, remote control, farm building, cedar chest, respirator, gas heater, meat counter, drilling platform, climbing bird's nest fern, sulphur butterfly, serger, butterfly bush, funnel web, sandal, police van, workspace, child's room, birthday cake, weightlifter, pretzel, cauliflower, fountain, recorder player, adam's peak, barbershop, basketball play, paving stone, trolleybus, obelisk, pig, swimsuit, northern pocket gopher, hypodermic syringe, titration, ctenophore family, snorkeling, refectory table, teapot, beer barrel, spectacled caiman, maypole, cummerbund, lorgnette, ice lolly, organ stop, ski jump, croquet mallet, telephone pole, birdhouse, dirndl, picket fence, dall sheep, sheepskin coat, tree cricket, brown bear, seller's market, guacamole, wood vise, guinea pig, thatch, inscription, western blind snake, cymbalist, chasuble, military officer, pottery, digital watch, two-spotted ladybug, jewelled headdress, miniskirt, teddy, milking shorthorn, bed jacket, trilobite, beer bottle, lake champlain, ice hockey rink, neck brace, stairway, ibizan hound, silky terrier, rain barrel, backpack, wandering albatross, american bison, crab nebula, binoculars, dustpan, great-nephew, matzo ball, ventriloquism, oort cloud, sushi bar, shofar, fire-eater, brown snail, macadamia integrifolia, genus andrena, monaco-ville, koala, variegated horsetail.