Language-driven Fine-grained Retrieval

Shijie Wang[1], Xin Yu[2], Yadan Luo[1], Zijian Wang[1], Pengfei Zhang[1], Zi Huang[1]
[1] The University of Queensland, Australia [2] The University of Adelaide, Australia

Abstract

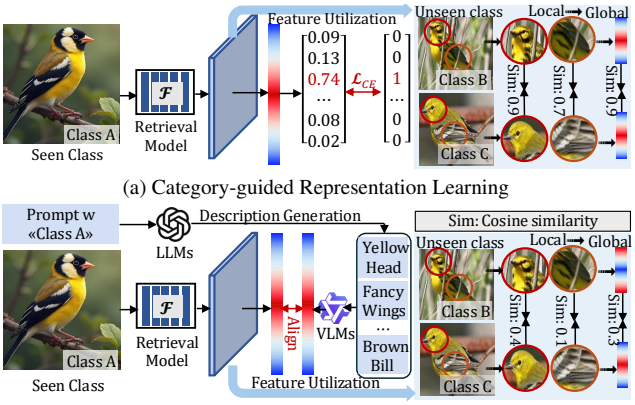
*Existing fine-grained image retrieval (FGIR) methods learn discriminative embeddings by adopting semantically sparse one-hot labels derived from category names as supervision. While effective on seen classes, such supervision overlooks the rich semantics encoded in category names, hindering the modeling of comparability among cross-category details and, in turn, limiting generalization to unseen categories. To tackle this, we introduce LaFG, a **La**nguage-driven framework for **F**ine-**G**rained Retrieval that converts class names into attribute-level supervision using large language models (LLMs) and vision–language models (VLMs). Treating each name as a semantic anchor, LaFG prompts an LLM to generate detailed, attribute-oriented descriptions. To mitigate attribute omission in these descriptions, it leverages a frozen VLM to project them into a vision-aligned space, clustering them into a dataset-wide attribute vocabulary while harvesting complementary attributes from related categories. Leveraging this vocabulary, a global prompt template selects category-relevant attributes, which are aggregated into category-specific linguistic prototypes. These prototypes supervise the retrieval model to steer it toward pinpointing visual details consistent with linguistic descriptions, thus modeling comparability among object details. Extensive evaluations show that LaFG achieves impressive performance on both fine- and coarse-grained benchmarks and generalizes well to unseen categories.*

1. Introduction

Fine-grained image retrieval (FGIR) aims to retrieve visually similar images from the same subcategory, even for categories unseen during training. This capability is pivotal to real-world applications, such as fashion recommendation (e.g., fine-grained clothing retrieval [1, 44, 45, 51, 57]) and ecological monitoring (e.g., endangered-species recognition [53, 54, 56, 58]). Motivated by its practical value, a large body of research has focused on learning representations that are both discriminative and generalizable to advance FGIR performance.

Recent studies on FGIR [20, 30, 32] rely on supervisory signals derived from one-hot encodings of category names to retrieve visually similar subcategories. While effective on seen categories, such semantically sparse supervision fails to model comparability among cross-category details, hindering the acquisition of transferable discriminative details and ultimately limiting generalization to unseen categories. As illustrated in Fig. 1a, visually similar parts (*e.g.*, head or wing patterns) across unseen classes remain difficult to distinguish since one-hot supervision lacks part-aware guidance. Consequently, their local features collapse into similar representations, resulting in highly similar retrieval embeddings. This gap suggests a novel investigation: rather than compressing category information into one-hot labels, can we unlock the semantics carried by category names and use them as supervisory signals that clearly reflects the comparability of object details?

This investigation aligns with vision–language models (VLMs) [2, 12, 31] that unlock rich semantic priors from



(a) Category-guided Representation Learning

(b) Language-driven Representation Learning

Figure 1. Motivation of LaFG. (a) Learning with one-hot labels compresses class names into a single global identifiers and overlooks parts and attributes, making it hard to compare appearance details when facing unseen categories. Hence, similar local regions become indistinguishable, which degrades generalization to unseen categories. (b) Language-driven learning turns category names into linguistic supervision, thus establishing detail comparability. The model acquires transferable discriminative knowledge and improves retrieval on unseen categories.

category names, naturally motivating language-driven supervision. However, prevailing methods typically conceptualize category names as global identifiers, aligning entire images to them while consequently overlooking the comparability of object details. Fortunately, recent advances in large language models (LLMs) [10, 26] make it feasible to generate detailed descriptions of object properties at scale, prompted by category names. Hence, we map LLM-generated descriptions into a vision-aligned embedding space via a frozen VLM and use these embeddings to supervise the retrieval model, thereby establishing detail comparability and facilitating generalization to unseen categories, as shown in Fig. 1b. Yet, raw LLM outputs are often incomplete, redundant, or noisy, making them inadequate for representing object appearances and thus providing unreliable supervision [11, 21]. The key challenge, therefore, lies in designing a robust framework that automatically generates, refines, and aligns textual semantics with visual evidence, expanding category names into a more expressive and effective supervisory signal.

To this end, we introduce LaFG, a language-driven framework for fine-grained retrieval that redefines a category name not as an index but as a semantic anchor, moving beyond the semantic narrowness of one-hot labels. We first prompt an LLM to unfold each name into diverse descriptions covering fine-grained attributes. These descriptions are mapped into a vision-aligned semantic space by a frozen VLM, and their embeddings are clustered across classes to induce a compact attribute vocabulary spanning the dataset. The vocabulary serves two roles: denoising and completion—removing redundancy and noise in raw LLM outputs while borrowing complementary attributes from visually related categories. With a global prompt template, we query this vocabulary to retrieve a sparse attribute set and aggregate it into a class-specific linguistic prototype, yielding attribute-level targets that replace one-hot labels. Finally, these prototypes supervise the retrieval model to steer it toward pinpointing visual details consistent with linguistic descriptions within these prototypes, thus modeling comparability among object details.

Our main contributions are summarized below:

- To the best of our knowledge, we are the first to expand category names beyond one-hot labels into semantically rich supervisory signals, thereby improving fine-grained generalization to unseen categories.
- We propose LaFG, a novel framework that implicitly model comparability between subtle object details via coupling LLMs and VLMs to automatically generate, refine, and align textual semantics with visual evidence.
- Extensive experiments show that our LaFG achieves state-of-the-art performance on widely-used fine-grained and coarse-grained retrieval benchmarks.

2. Related Works

Fine-grained Image Retrieval. Recent advances in fine-grained image retrieval have primarily evolved along two distinct technical pathways [16, 18, 27, 33, 46, 47, 49, 50, 52, 55, 64–66]. Localization-based methods, exemplified by works like A^2-Net [58], focus on precise object or part localization to enhance retrieval accuracy through reconstruction-based learning. Concurrently, metric-based approaches such as DDML [30] attempt to learn discriminative embedding spaces through sophisticated distance metrics, while NIA [34] enforces unique translatability from class proxies, pulling same-subcategory samples closer in the embedding space. However, such methods, which rely on supervisory signals derived from one-hot encodings of category names, struggles to capture comparability among cross-category details due to semantically sparse supervision, thereby limiting generalization to unseen categories. To address this, our work, LaFG, extends category names beyond one-hot labels into semantically enriched supervisory signals, implicitly modeling comparability among object details and acquiring transferable discriminative details.

Vision-language Alignment. Vision–language alignment [13, 23, 31, 39, 48] learns powerful joint representations across modalities by pre-training on large-scale image–text pairs. To mitigate modality discrepancy, existing approaches refine contrastive objectives to achieve more precise cross-modal alignment, either from the perspective of token-level correspondence [3, 43, 61] or multi-level semantic consistency [22, 24, 62]. Hence, these works provide more accurate and richer supervision signals of multiple granularity, thus semantically aligning vision and language. Unlike prior methods, we project visual embeddings into the linguistic prototype space induced by the VLM through distribution-wise maximum similarity learning. In this way, LaFG steers the retrieval model toward pinpointing visual details consistent with linguistic descriptions within these prototypes, while simultaneously preserving instance-specific cues from the visual inputs.

Language-guided Learning. Language-guided learning has been widely applied across diverse domains, including global image editing [37], single-image reflection separation [67], visual continual learning [28], and semantic segmentation [9]. They typically employ language modeling as a pretext task for visual learning [5, 35], using linguistic supervision from pre-trained language models to guide the learning of visual representations. However, most existing methods adopt a unidirectional paradigm using static linguistic features as fixed supervision for visual extraction, but fail to account for scenarios where language descriptions are imprecise or incomplete. In contrast, we construct an attribute vocabulary from LLM-generated descriptions and exploit the VLM's shared vision–language space for noise-robust and complementary attribute selection.

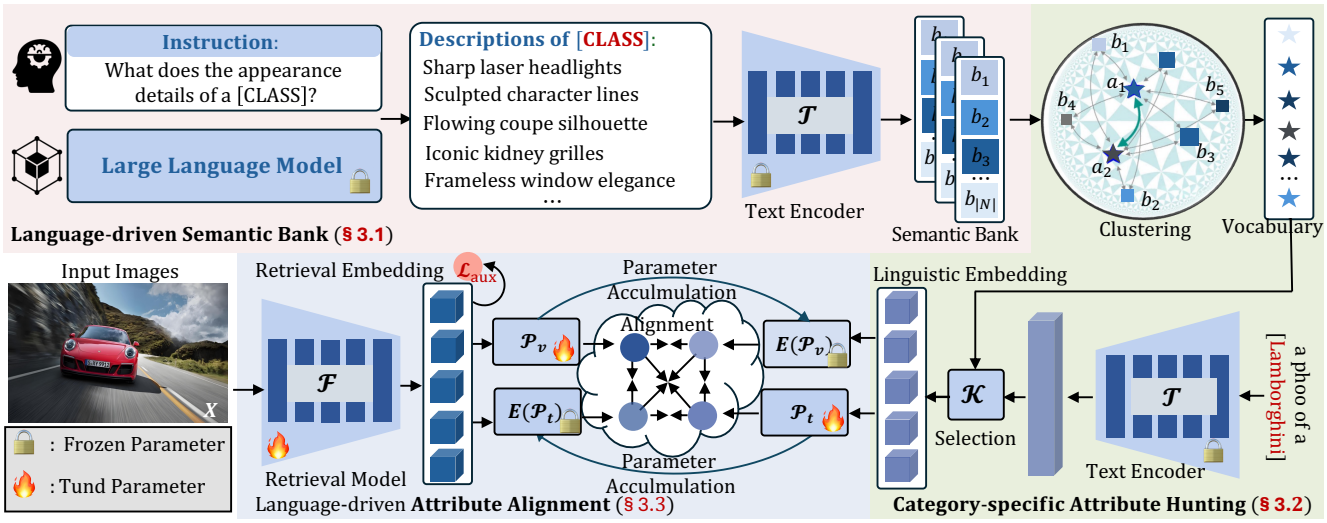


Figure 2. Framework illustration of **Language-driven Fine-grained Generalization**. See §3 for more details.

3. Methodology

The core of LaFG (Fig. 2) is to convert category names into semantically rich supervision. Given a category name, an LLM generates attribute-oriented descriptions, which a frozen VLM encodes to form a semantic bank. Cross-class clustering produces a compact, shared attribute vocabulary. A global prompt then selects a sparse, discriminative attribute set per category and aggregates it into a linguistic prototype that replaces one-hot labels. Finally, distribution alignment ties these prototypes to visual features, preserving instance-specific cues while enforcing language-consistent appearance modeling.

3.1. Language-driven Semantic Bank

One-hot labels provide limited semantic supervision, weakening cross-category fine-grained comparability. We address this by treating category names as semantic anchors, prompting an LLM to produce attribute-oriented descriptions, and encoding them with a VLM into a vision-aligned embedding space, thereby bridging the vision–language modality gap and forming a comprehensive semantic bank.

To align generic LLM knowledge with fine-grained visual semantics, we enrich subcategory names with fine-grained instructions. Specifically, we prompt the frozen LLM (*e.g.*, GPT-4 [29]) as follows:

> "Generate n distinct and descriptive statements that capture the key visual attributes of `[CLASS]`. Include holistic semantic characteristics and fine-grained textural details that would help distinguish `[CLASS]` from other visually similar subcategories. Texts should be of the form: an appearance description of `[CLASS]`. It + descriptive contexts".

Guided by this prompt, we obtain n descriptions $\mathrm{D_c}$ per class c, each encoding rich visual semantics.

Next, we build a comprehensive semantic bank by leveraging a frozen pre-trained VLM text encoder $\Phi_{\mathrm{t}}(\cdot)$ (*e.g.*, CLIP [31]) to embed the generated descriptions into a unified semantic space. This encoder possesses the unique ability to interpret language from a visual perspective, thereby mapping textual descriptions onto a vision-aligned feature manifold. Specifically, for each class c, we encode its n descriptions $\mathrm{D_c}$ into a set of attribute embeddings:

$$\mathrm{B_c} = \left\{ \Phi_{\mathrm{t}}(\mathrm{D_c^i}) \mid \mathrm{D_c^i} \in \mathrm{D_c} \right\}, \tag{1}$$

where $\mathrm{B_c} \in \mathbb{R}^{\mathrm{n}\times\mathrm{d}}$ represents the set of n attribute embeddings with the dimension d for class c. This bank, denoted as $\mathcal{B} = [\mathrm{B_1}, \cdots, \mathrm{B_C}] \in \mathbb{R}^{\mathrm{C}\times\mathrm{n}\times\mathrm{d}}$, serves as a semantic repository that encapsulates visual semantics across all training categories, where C denotes the number of classes in the training set, and d is the dimension of embeddings.

Unlike prior VLM-based approaches that directly rely on category names with limited fine-grained semantics, our method leverages the generative capability of LLMs to produce semantic-enriched descriptions, thereby enhancing generalization to unseen subcategories.

3.2. Category-aware Attribute Hunting

Leveraging these fine-grained semantics sourced from LLMs for representing fine-grained categories remains unreliable. Although LLMs can produce diverse descriptions under fine-grained prompts, they often miss portions of object appearances and introduce redundancy or noise with a high probability, primarily due to the lack of explicit fine-grained supervision during their pretraining.

Consequently, we refrain from directly fusing per-class texts from the semantic bank. Instead, we cluster all description embeddings across the training set to construct a global attribute vocabulary. The vocabulary serves two

roles: denoising and completion—removing redundancy and noise in raw LLM outputs while borrowing complementary semantics from visually related categories. Given this vocabulary, we then use a simple hand-crafted retrieval template to hunt, for each category, the most semantically relevant attributes, yielding a compact yet expressive attribute set tailored to that category.

We consolidate the language-driven semantic bank $\mathcal{B}$ into a compact attribute vocabulary set $\mathcal{V}$. Concretely, we group embeddings into $|\mathrm{N}|$ universal attributes by applying K-means clustering $\mathcal{K}(\cdot)$ across all training categories:

$$\mathcal{V} = \mathcal{K}(\mathcal{B}, |\mathrm{N}|) = \{\mathrm{a_i}\}_{\mathrm{i}=1}^{|\mathrm{N}|}, \tag{2}$$

where each cluster centroid $a_i \in \mathbb{R}^{\mathrm{d}}$ represents an attribute, a common semantic pattern that consistently emerge across multiple fine-grained descriptions. This clustering process automatically consolidates recurring semantic patterns and eliminates redundant descriptions, yielding a discriminative attribute vocabulary that effectively captures the essential visual traits shared across subcategories.

Building upon the universal attribute vocabulary $\mathcal{V}$, we perform category-aware attribute selection through a cross-text semantic alignment. For each class c, we generate a hand-craft descriptor "a photo of [CLASS]" and encode it through the same CLIP's text encoder $\Phi_{\mathrm{t}}(\cdot)$ to obtain the category embedding $t_c \in \mathbb{R}^d$. Leveraging CLIP's shared vision-language embedding space, the category embedding encapsulates category-level semantic centroids [31, 68]. We thus employ it as a query to retrieve the most semantically relevant attributes from $\mathcal{V}$ based on their similarities. The selection process identifies the Top-K attributes that maximize the similarity with the class prototype:

$$\mathcal{V}_{\mathrm{c}} = \{\mathrm{a_k} : \mathrm{k} \in \underset{\mathrm{Top-K}}{\arg\max}\{\mathrm{t_c^T} \cdot \mathrm{a_k}\}_{\mathrm{k}=1}^{\mathrm{B}}\}. \tag{3}$$

We compute the final category prototype $\mathrm{T_c} \in \mathbb{R}^{\mathrm{d}}$ for class c by adaptively fusing its class embedding t_c with the most relevant attributes $\mathcal{V}_c \in \mathbb{R}^{\mathrm{K}\times\mathrm{d}}$. The representation is formulated as:

$$\mathrm{T_c} = \mathrm{t_c} + \sum_{\mathrm{k}=1}^{\mathrm{K}} \sigma\big(\mathrm{t_c}^{\top}\mathrm{a_k}\big) \cdot \mathrm{a_k}, \tag{4}$$

where $\sigma(\cdot)$ applies softmax normalization to the similarity scores between $\mathrm{t_c}$ and each attribute $\mathrm{a_k} \in \mathcal{V}_{\mathrm{c}}$.

3.3. Language-driven Attribute Alignment

Aligning visual cues present in images with attribute-level representations derived from category names is a prerequisite for establishing comparability among fine-grained object details. To this end, we propose a language-driven attribute alignment module that supervise the retrieval model to steer it toward pinpointing visual details consistent with LLM-generated linguistic descriptions.

For an input image $\mathbf{X} \in \mathbb{R}^{3\times H\times W}$, the retrieval model $\mathcal{F}$ extracts its representation as an embedding vector $\mathrm{V} \in \mathbb{R}^{\mathrm{d}}$. Concurrently, we select the category prototype $\mathrm{T_c}$ corresponding to the image's category c. To supervise $\mathcal{F}$ towards learning visual features that align with linguistic descriptions, we first introduce two modality-specific linear projectors: $\mathcal{P}_{\mathrm{v}}$ for the visual embedding V and $\mathcal{P}_{\mathrm{t}}$ for the category-specific prototype $\mathrm{T_c}$, mapping their respective inputs into attribute-aligned embedding spaces.

Since each projector receives input from only one modality, it learns to model the attribute distribution specific to this modality. If both projectors produce the same distribution for a given embedding—regardless of its original modality—this indicates that the embedding has become modality-invariant, effectively supervise the retrieval model towards learning visual features that align with linguistic descriptions. This alignment is quantified by minimizing the symmetric Kullback-Leibler (KL) divergence

$$\begin{aligned}\hat{\mathcal{L}}_{\mathrm{ali}} = \mathcal{P}_{\mathrm{v}}(\mathrm{T_c}|\theta_{\mathrm{v}}) \log \frac{\mathcal{P}_{\mathrm{v}}(\mathrm{T_c}|\theta_{\mathrm{v}})}{\mathcal{P}_{\mathrm{t}}(\mathrm{T_c}|\theta_{\mathrm{t}})} \\ + \mathcal{P}_{\mathrm{t}}(\mathrm{V}|\theta_{\mathrm{t}}) \log \frac{\mathcal{P}_{\mathrm{t}}(\mathrm{V}|\theta_{\mathrm{t}})}{\mathcal{P}_{\mathrm{v}}(\mathrm{V}|\theta_{\mathrm{v}})},\end{aligned} \tag{5}$$

where θ_{t} and θ_{v} denotes the parameters of $\mathcal{P}_{\mathrm{t}}$ and $\mathcal{P}_{\mathrm{v}}$, respectively.

However, we observe that direct optimization of this objective leads to premature convergence, where the projectors simply mimic each other's outputs without establishing meaningful attribute distribution alignment. To address this, we introduce mean projectors with exponentially moving average parameters that stabilize the training process:

$$\begin{aligned}\mathrm{E}^{(\mathrm{t})}[\theta_{\mathrm{v}}] &= (1-\alpha)\,\mathrm{E}^{(\mathrm{t}-1)}[\theta_{\mathrm{v}}] + \alpha\,\theta_{\mathrm{v}}, \\ \mathrm{E}^{(\mathrm{t})}[\theta_{\mathrm{t}}] &= (1-\alpha)\,\mathrm{E}^{(\mathrm{t}-1)}[\theta_{\mathrm{t}}] + \alpha\,\theta_{\mathrm{t}},\end{aligned} \tag{6}$$

where $\mathrm{E}^{(\mathrm{t})}[\theta]$ and $\mathrm{E}^{(\mathrm{t}-1)}[\theta]$ denote the parameters of mean projectors in the current iteration t and last iteration $\mathrm{t}-1$. The mean projectors are initialized as $\mathrm{E}^{(0)}[\theta_{\mathrm{v}}] = \theta_{\mathrm{v}}$ and $\mathrm{E}^{(0)}[\theta_{\mathrm{t}}] = \theta_{\mathrm{t}}$. The parameter α is the updating ratio within the range of $(0, 1]$.

In this way, Eq. (5) can be rewritten as

$$\begin{aligned}\mathcal{L}_{\mathrm{ali}} = \mathrm{P_v}(\mathrm{T_c}|\mathrm{E}[\theta_{\mathrm{v}}])\log \frac{\mathrm{P_v}(\mathrm{T_c}|\mathrm{E}[\theta_{\mathrm{v}}])}{\mathrm{P_t}(\mathrm{T_c}|\theta_{\mathrm{v}})} \\ + \mathrm{P_t}(\mathrm{V}|\mathrm{E}[\theta_{\mathrm{t}}]) \log \frac{\mathrm{P_t}(\mathrm{V}|[\theta_{\mathrm{t}}])}{\mathrm{P_t}(\mathrm{V}|\theta_{\mathrm{v}})}.\end{aligned} \tag{7}$$

During training, since the two projectors are not updated via backpropagation, $\mathcal{L}_{\mathrm{ali}}$ exclusively optimizes the retrieval model by enforcing the distribution of visual embeddings

to align with the corresponding category prototype. This alignment allows linguistic attributes to attend to diverse visual regions within images, effectively modeling comparability among fine-grained details and enhancing generalization to unseen categories. Crucially, LaFG aligns distributions instead of than individual embeddings, allowing retrieval embeddings to preserve instance-specific cues and stay consistent with linguistic descriptions.

3.4. Overall Training Objective

By transforming category names into semantically rich prototypes, the retrieval model is supervised to establish comparability among fine-grained visual details associated with each category. To extend this comparability beyond individual instances, we introduce an auxiliary contrastive loss that encouraging the retrieval model to learn comparability among cross-category details.

During training, we sample N categories with two instances per class, resulting in a batch size of $\mathrm{K} = 2\mathrm{N}$. For an anchor $\mathrm{z_i}$, the auxiliary contrastive loss is defined as:

$$\mathcal{L}_{\mathrm{aux}}(\mathrm{z_i}) = -\log \frac{\exp\left(-\mathrm{D}(\mathrm{z_i}, \mathrm{z_j})/\tau\right)}{\sum_{\mathrm{k}=1,\mathrm{k}\neq\mathrm{i}}^{\mathrm{K}} \exp\left(-\mathrm{D}(\mathrm{z_i}, \mathrm{z_k})/\tau\right)}, \quad (8)$$

where $(\mathrm{z_i}, \mathrm{z_j})$ denotes a positive pair from the same subcategory, and $(\mathrm{z_i}, \mathrm{z_k})$ represents all other sample pairs except the anchor itself. Here, τ is a temperature hyperparameter, and $\mathrm{D}(\cdot, \cdot)$ represents the distance, implemented as the squared Euclidean distance between normalized vectors:

$$\mathrm{D}(\mathrm{z_i}, \mathrm{z_j}) = \left\| \frac{\mathrm{z_i}}{\|\mathrm{z_i}\|_2} - \frac{\mathrm{z_j}}{\|\mathrm{z_j}\|_2} \right\|_2^2 = 2 - 2\frac{\langle \mathrm{z_i}, \mathrm{z_j} \rangle}{\|\mathrm{z_i}\|_2 \cdot \|\mathrm{z_j}\|_2}. \quad (9)$$

The total loss $\mathcal{L}$ of LaFG is formulated as

$$\mathcal{L} = \mathcal{L}_{\mathrm{aux}} + \beta \cdot \mathcal{L}_{\mathrm{ali}}, \quad (10)$$

where β is the hyper-parameter to balance the contributions of the individual loss item.

4. Experiments

4.1. Experimental Setup

Datasets. CUB-200-2011 [4] consists of 200 bird species. We use the first 100 subcategories (5,864 images) for training and the consists of (5,924 images) for testing. The Stanford Cars [19] includes 196 car models. Similarly, we use the first 98 classes, which contain 8,054 images, for training and the remaining classes, which contain 8,131 images, for testing. Stanford Online Products (SOP) [36] is divided into the 11, 318 subcategories (59, 551 images) in training, and the rest 11, 316 classes (60, 502 images) in testing. *This split ensures **no category overlap** between training and testing sets, where all testing categories are strictly unseen during training to evaluate cross-category generalization.*

Table 1. Recall@1 performance comparison across constraint combinations on CUB-200-2011 benchmark.

$\mathcal{L}_{aux}$	$\hat{\mathcal{L}}_{ali}$	$\mathcal{L}_{ali}$	Recall@1
✓			82.6%
✓	✓		85.3%$_{+2.7\%}$
		✓	86.5%$_{+3.9\%}$
✓		✓	**87.2%**$_{+4.2\%}$

Table 2. Evaluation results of retrieval performance on CUB-200-2011 with/without the synergy of LLMs and VLMs. "Handcrafted Language" indicates the use of template-based textual descriptions (e.g., a photo of a [·]) instead of LLM-generated descriptions.

Language	Recall@1
VLM + Hand-craft Language	83.7%
VLM + LLM (w/o Vocabulary)	85.3%$_{+2.6\%}$
VLM + LLM	**87.2%**$_{+3.5\%}$

Implementation Details. Our retrieval model is implemented on top of a Vision Transformer (ViT) [6] initialized with ImageNet pre-trained weights. Input images are resized to 256×256 and randomly cropped to 224×224 during training. We adopt stochastic gradient descent with an initial learning rate of 1×10^{-5}, momentum of 0.9, and weight decay of 1×10^{-4}, using a batch size of 900 on NVIDIA A100 GPUs. To improve robustness, we apply standard data augmentations including random cropping, horizontal flipping, and color jittering. The learning rate follows an exponential decay schedule with a decay factor of 0.9 every five epochs over a total of 200 training epochs.

Evaluation protocols. We evaluate retrieval performance using Recall@K with cosine distance, following the standard protocol in prior work [38]. Specifically, for each query image, the model retrieves the top-M most similar images. A score of 1 is assigned if at least one positive image appears within the top-M results, and 0 otherwise. The final Recall@K is obtained by averaging these scores across all queries in the test set.

4.2. Ablation Experiments

Efficacy of various constraints. LaFG is optimized with a combination of two objectives: an auxiliary loss $\mathcal{L}_{\mathrm{aux}}$ and an alignment loss $\mathcal{L}_{\mathrm{ali}}$ (or its variant $\hat{\mathcal{L}}_{\mathrm{ali}}$). These losses serve complementary roles to effectively convert category names into semantically rich supervision, thus establishing the comparability of cross-category details. Ablation studies on CUB-200-2011 (Tab. 1) reveal critical insights: Training solely with $\mathcal{L}_{\mathrm{aux}}$ achieves 82.6% Recall@1 but fails to model detail comparability, limiting generaliza-

Table 3. Compared with competitive methods on CUB-200-2011, Stanford Cars 196 and Stanford Online Products datasets. "Arch" denotes the backbone architecture. "R50" and "ViT" denote the ResNet-50 [8] and Vision Transformer [6] backbones, respectively.

Method	Arch	CUB-200-2011				Stanford Cars 196				Stanford Online Products			
		1	2	4	8	1	2	4	8	1	10	100	1000
CBML $_{\text{TPAMI23}}$ [15]	R50	69.9	80.4	87.2	92.5	88.1	92.6	95.4	97.4	79.9	91.5	96.5	98.9
NIR$_{\text{CVPR22}}$ [34]	R50	70.5	80.6	-	-	89.1	93.4	-	-	80.4	91.4	-	-
HSE$_{\text{ICCV23}}$ [60]	R50	70.6	80.1	87.1	-	89.6	93.8	96.0	-	80.0	91.4	96.3	-
IDML$_{\text{TPAMI24}}$ [42]	R50	70.7	80.2	-	-	90.6	94.5	-	81.5	-	-	-	-
HIST $_{\text{CVPR22}}$ [25]	R50	71.4	81.1	88.1	-	89.6	93.9	96.4	-	81.4	92.0	96.7	-
PNCA++$_{\text{ECCV20}}$ [40]	R50	72.2	82.0	89.2	93.5	90.1	94.5	97.0	98.4	81.4	92.4	96.9	99.0
DIML$_{\text{TPAMI24}}$ [63]	ViT	76.7	-	-	-	80.7	-	-	-	79.5	-	-	-
DFML-PA$_{\text{CVPR23}}$ [41]	ViT	79.1	86.8	-	-	89.5	93.9	-	-	84.2	93.8	-	-
DVA $_{\text{AAAI26}}$ [14]	ViT	84.9	90.6	94.5	96.7	90.7	94.8	97.1	98.4	-	-	-	-
DPHM$_{\text{PR25}}$ [59]	ViT	85.5	91.3	94.6	96.6	84.1	90.5	94.5	97.1	84.8	94.5	98.1	99.4
HypViT$_{\text{CVPR22}}$ [7]	ViT	85.6	91.4	94.8	96.7	89.2	94.1	96.7	98.1	85.9	94.9	98.1	99.5
HIER$_{\text{CVPR23}}$ [17]	ViT	85.7	91.3	94.4	-	88.3	93.2	96.1	-	86.1	95.0	98.0	-
SEE$_{\text{IJCAI25}}$ [20]	ViT	85.8	91.4	94.6	-	88.8	93.8	96.4	-	86.3	95.0	98.2	-
DDML$_{\text{AAAI25}}$ [30]	ViT	86.0	91.7	95.2	96.8	89.5	94.2	96.8	98.2	86.1	95.1	98.2	99.5
VPTSP-G$_{\text{ICLR24}}$ [32]	ViT	86.6	91.7	94.8	-	87.7	93.3	96.1	-	84.4	93.6	97.3	-
Our LaFG	ViT	**87.2**	**92.4**	**95.2**	**97.0**	**91.5**	**94.6**	**96.6**	**98.5**	**87.1**	**95.8**	**98.5**	**99.5**

tion to unseen categories. Incorporating $\hat{\mathcal{L}}_{\text{ali}}$ boosts Recall@1 to 85.3%, demonstrating that LLM-generated descriptions provide valuable signals. However, by optimizing projectors to mimic the distributions of category-specific prototypes rather than directly refining visual representations, $\hat{\mathcal{L}}_{\text{ali}}$ remains insufficient. To overcome this, we introduce exponential moving average (EMA) updates replacing backpropagation. This innovation enables $\mathcal{L}_{\text{ali}}$ to directly optimize the retrieval model sensitive to semantic comparison. Upon this, the group of $\mathcal{L}_{\text{ali}}$ and $\mathcal{L}_{\text{aux}}$ successfully establishes comparability among cross-category details, thus acquiring state-of-the-art 87.2% Recall@1 on the CUB-200-2011 benchmark.

Importance of the synergy of LLMs and VLMs. We assess the synergy between LLMs and VLMs by varying both the source and the processing of textual descriptions (Tab. 2). Using handcrafted CLIP prompts ("a photo of a [·]") as a baseline yields only marginal gains, since such templates lack the fine-grained cues contributing to decision boundary. Replacing them with LLM-generated, attribute-centric descriptions injects complementary semantics and raises Recall@1 to 85.3%. However, raw LLM texts remain incomplete and noisy. To consolidate and share attributes across related subcategories, we cluster VLM text embeddings of all descriptions to induce a compact, dataset-wide attribute vocabulary, then retrieve a relevant subset per category to form a category-specific prototype. These prototypes supervise the retrieval model to attend to attribute-consistent visual details, boosting accuracy to 87.2% and underscoring the effectiveness of coupling LLMs with VLMs.

4.3. Comparison with State-of-the-art Methods

Fine-grained Image Retrieval. Our LaFG establishes new state-of-the-art performance on both zero-shot fine-grained image retrieval benchmarks (CUB-200-2011 and Stanford Cars-196), achieving Recall@1 accuracies of 87.2% and 91.5% respectively, as detailed in Tab. 3. We achieve absolute Recall@1 gains over SEE [20] (+1.4% on CUB, +2.7% on Cars) and DDML [30] (+1.2% on CUB, +2.0% on Cars), highlighting the advantage of converting category names from one-hot labels to seamntically rich supervisory signals. Furthermore, prior work such as VPTSP-G [32] typically exploits the zero-shot capability of foundation models to learn discriminative and generalizable embeddings. In contrast, LaFG employs a frozen VLM to project LLM-generated descriptions into a vision-aligned semantic space, induces category-specific attribute prototypes, and uses these prototypes to supervise the retrieval model, guiding it to localize visual details consistent with language descriptions and yielding superior performance. The consistent improvements substantiate that converting class names into attribute-level supervision explicitly models cross-category comparability, thereby enhancing generalization to unseen categories.

Coarse-grained Image Retrieval. To assess LaFG's generalization capacity beyond fine-grained domains, we conduct a large-scale evaluation on the coarse-grained

Table 4. Retrieval performance on CUB-200-2011 for models trained with different numbers of language descriptions n generated by LLMs.

Number n	5	10	15	20	25
Recall@1	81.4%	84.3%	85.7%	87.2%	86.3%

Table 5. Retrieval performance on CUB-200-2011 for models trained with different attribute vocabulary sizes $|\mathrm{N}|$.

Size $\|\mathrm{N}\|$	32	64	128	256	512
Recall@1	84.3%	85.7%	87.2%	86.3%	85.9%

Table 6. Evaluation on CUB-200-2011 for models trained with different Top-K category-specific attributes used to represent each training category in Eq. (3).

Top-K	16	24	32	40	48
Recall@1	84.1%	85.2%	86.3%	87.2%	86.9%

Table 7. Quantitative performance of the model on CUB-200-2011 when trained with different dimension of distribution in the two projectors.

Dimension	32	64	128	256	512
Recall@1	80.1%	84.5%	86.0%	87.2%	86.9%

Table 8. Evaluation on CUB-200-2011 for models trained with different update ratios α in the parameter updating defined in Eq. (6).

Ratio α	0.1	0.2	0.4	0.6	0.8
Recall@1	87.1%	87.2%	86.9%	86.6%	86.6%

Table 9. Quantitative performance of the model on CUB-200-2011 when trained with different weights β in the loss function defined in Eq. (10).

Weight β	1	5	10	15	20
Recall@1	83.5%	86.3%	87.2%	87.1%	86.5%

benchmark, *i.e.*, Stanford Online Products, as reported in Tab. 3. Given the absence of fine-grained subcategory names in SOP, we leverage coarse labels to guide attribute generation and replace textual queries with image-derived embeddings to facilitate the attribute selection process. The framework synergistically leverages LLMs and VLMs to project LLM-generated descriptions into a vision-aligned space via VLMs, enabling accurate representation of both seen and unseen categories. This synergy enables LaFG to resolve subtle fine-grained distinctions and to bridge coarse-grained semantic gaps, yielding superior performance across levels of granularity. Consistent improvements on SOP further confirm that converting class names into attribute-level supervision scales effectively from fine-grained to coarse-grained retrieval.

4.4. Discussion

Effect on different numbers of language descriptions. As shown in Tab. 4, retrieval performance on CUB-200-2011 exhibits a clear dependence on the number of language descriptions (n) generated by LLMs. Recall@1 steadily increases with larger n, peaking at 87.2% when $n = 20$, before slightly decreasing to 86.3% at $n = 25$. This non-monotonic behavior reveals two competing effects: when $n \leq 10$, insufficient linguistic cues fail to capture fine-grained visual distinctions, resulting in suboptimal performance (81.4%–84.3%); when $n > 20$, redundant or noisy descriptions begin to obscure discriminative attributes, causing a 1.1% drop from $n = 20$ to $n = 25$. Overall, using around 20 high-quality descriptions achieves the best trade-off between semantic richness and precision for fine-grained representation learning.

Investigation on the size of attribute vocabulary. The results in Tab. 5 demonstrate a non-linear relationship between attribute vocabulary size ($|\mathrm{N}|$) and retrieval performance. Recall@1 peaks at 87.2% when $|\mathrm{N}| = 128$, while both smaller and larger vocabularies yield reduced effectiveness. With a small vocabulary size (e.g., $|\mathrm{N}| = 32$), multiple distinctive attributes collapse into a single cluster, leading to the loss of discriminative details. In contrast, an excessively large vocabulary (e.g., $|\mathrm{N}| = 512$) causes over-segmentation, where semantically similar descriptions are divided into separate attributes, introducing redundancy and noise that degrade performance. The optimal $|\mathrm{N}| = 128$ strikes an effective balance between preserving fine-grained distinctiveness and maintaining attribute robustness.

How many attributes effectively represent objects? Tab. 6 presents the impact of Top-K attribute selection on retrieval performance. Recall@1 increases with larger K, peaking at 87.2% when $\mathrm{K} = 40$. Specifically, when using a smaller K, the limited number of attributes fails to capture the full range of visual cues within each category, resulting in incomplete semantic representation. In contrast, larger K values introduce irrelevant attributes from other categories, reducing discriminative power. The optimal setting ($\mathrm{K} = 40$) strikes the best trade-off, capturing meaningful attributes while remaining selective enough to preserve inter-class distinctiveness. The minor 0.3% drop at $\mathrm{K} = 48$ indicates slight attribute confusion, emphasizing the need for careful attribute selection in fine-grained retrieval.

Effect on dimension of attribution distribution. In Tab. 7, retrieval accuracy improves as the prototype distribution di-

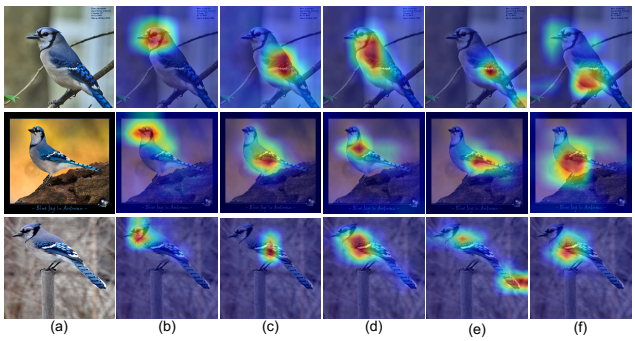


Figure 3. Visualization of clustered attribute responses for the same subcategory (e.g., Blue Jay). The Top-5 attributes are selected from the vocabulary based on their similarity scores. (a) Input images; (b)–(f) Attribute response regions for comparison.

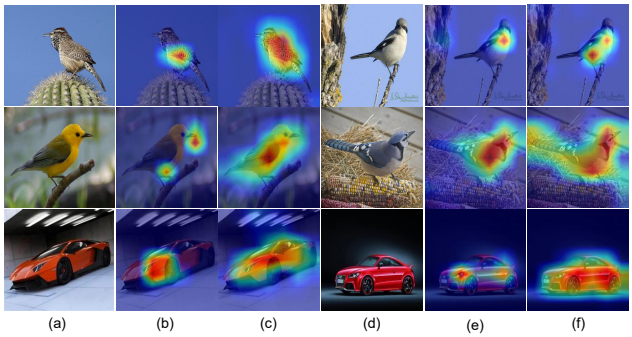


Figure 4. Illustration of class activation maps generated by the baseline and our LaFG. (a) and (d) show the input images; (b) and (e) present the corresponding class activation maps produced by the baseline; (c) and (f) display the maps generated by our LaFG.

mension increases from 32 to 256, indicating that higher capacity better captures the rich, fine-grained attributes needed to highlight subtle inter-class differences. However, pushing the dimension further to 512 slightly degrades performance, likely because an over-parameterized prototype space encourages visual features to overfit class-specific prototypes and overlook instance-specific cues. In contrast, a small dimension under-represents attribute diversity and limits the model's ability to obtain important distinctions.

Effect on parameter accumulation. Tab. 8 illustrates the effect of the exponentially moving average (EMA) update ratio (α) on model performance, highlighting key insights into parameter optimization dynamics. The best performance (87.2% Recall@1) at $\alpha = 0.2$ indicates an optimal balance between parameter stability, which preserves sufficient historical information, and model adaptability, which incorporates new discriminative cues. A smaller ratio ($\alpha = 0.1$) yields slightly lower performance due to over-reliance on historical parameters, which slows the integration of new attribute representations. Conversely, higher ratios ($\alpha \geq 0.4$) bias updates toward recent parameters, causing unstable training and noisy attribute projections.

Hyperparameter analysis. In Tab. 9, the loss weight β critically governs how effectively the model exploits implicit attribute supervision and enforces cross-category attribute comparability. Performance peaks at β=10, indicating a balanced interplay between the language-driven alignment loss and the auxiliary contrastive loss. Underweighting weakens attention to language-guided visual cues, whereas overweighting over-regularizes toward linguistic descriptions, thus degrading FGIR performance.

Visual attribute analysis. To evaluate whether the constructed attribute vocabulary effectively captures discriminative object characteristics, we visualize the top-5 clustered attributes ranked by similarity scores (Fig. 3), using the Blue Jay category as an example. Two key observations emerge: each attribute consistently highlights salient regions (e.g., head, neck, wings), indicating that the learned vocabulary aligns with coherent visual semantics. Moreover, different attributes focus on diverse visual cues (e.g., beak shape vs. wing pattern), demonstrating their complementary roles in fine-grained characterization. These findings confirm that LaFG encodes semantically meaningful attributes that correspond to localized visual features, provide diverse yet coherent appearance descriptors.

Feature representation analysis. As shown in Fig. 4, the class activation maps demonstrate how language descriptions enhance visual representation learning. Compared with the baseline, which narrowly focuses on the most discriminative regions, LaFG exhibits two notable advantages. First, it activates additional informative regions (e.g., subtle texture patterns) that the baseline overlooks, preserving features essential for recognizing unseen categories. Second, language guidance directs the model to attend to fine-grained appearance details (e.g., feather patterns) rather than coarse category-level semantics, thereby improving generalization. These visualizations demonstrate that language-guided training supplies complementary visual cues absent in conventional works, thereby expanding and refining the model's visual understanding and explaining LaFG's superior generalization performance.

5. Conclusion

This paper introduces LaFG, a language-driven framework for FGIR that redefines a category name not as an index but as a semantic anchor, moving beyond the semantic narrowness of one-hot labels. LaFG implicitly models comparability among object details and acquires transferable discriminative details, thus improving generalization to unseen categories. Comprehensive evaluations on fine-grained and coarse-grained benchmarks confirm that turning class names beyond one-hot labels into attribute-level supervisory signals provides large gains and strong generalization across seen and unseen categories.

Acknowledgements. This work was partially supported by ARC DE240100105, DP240101814, DP230101196, IE250100108, and ARC Industrial Transformation Research Hubs IH230100013.

References

[1] Kenan E. Ak, Ashraf A. Kassim, Joo-Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7708–7717. Computer Vision Foundation / IEEE Computer Society, 2018. 1

[2] Yassir Bendou, Amine Ouasfi, Vincent Gripon, and Adnane Boukhayma. Proker: A kernel perspective on few-shot adaptation of large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 25092–25102. Computer Vision Foundation / IEEE, 2025. 1

[3] Junyu Bi, Daixuan Cheng, Ping Yao, Bochen Pang, Yuefeng Zhan, Chuanguang Yang, Yujing Wang, Hao Sun, Weiwei Deng, and Qi Zhang. Vl-match: Enhancing vision-language pretraining with token-level and instance-level matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2584–2593, 2023. 2

[4] Steve Branson, Grant Van Horn, Serge J. Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *CoRR*, abs/1406.2952, 2014. 5

[5] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11162–11173. Computer Vision Foundation / IEEE, 2021. 2

[6] 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. 5, 6

[7] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan V. Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7399–7409. IEEE, 2022. 6

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016. 6

[9] Wenbin He, Suphanut Jamonnak, Liang Gou, and Liu Ren. CLIP-S4: language-guided self-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11207–11216. IEEE, 2023. 2

[10] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1049–1065. Association for Computational Linguistics, 2023. 2

[11] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55, 2025. 2

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 4904–4916. PMLR, 2021. 1

[13] 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*, pages 4904–4916. PMLR, 2021. 2

[14] Xin Jiang, Meiqi Cao, Hao Tang, Fei Shen, and Zechao Li. Fine-grained image retrieval via dual-vision adaptation. *arXiv preprint arXiv:2506.16273*, 2025. 6

[15] Shichao Kan, Zhiquan He, Yigang Cen, Yang Li, Vladimir Mladenovic, and Zhihai He. Contrastive bayesian analysis for deep metric learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6):7220–7238, 2023. 6

[16] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Embedding transfer with label relaxation for improved metric learning. In *CVPR*, pages 3967–3976. Computer Vision Foundation / IEEE, 2021. 2

[17] Sungyeon Kim, Boseung Jeong, and Suha Kwak. HIER: metric learning beyond class labels via hierarchical regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19903–19912. IEEE, 2023. 6

[18] ByungSoo Ko, Geonmo Gu, Han-Gyu Kim, and ByungSoo Ko. Learning with memory-based virtual classes for deep metric learning. In *ICCV*, pages 11772–11781. IEEE, 2021. 2

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561, 2013. 5

[20] Binh Minh Le and Simon S. Woo. SEE: spherical embedding expansion for improving deep metric learning (extended abstract). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 10906–10911. ijcai.org, 2025. 1, 6

[21] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 10879–10899. Association for Computational Linguistics, 2024. 2

[22] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, pages 121–137. Springer, 2020. 2

[23] 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*, 2021. 2

[24] Zejun Li, Zhihao Fan, Huaixiao Tou, Jingjing Chen, Zhongyu Wei, and Xuanjing Huang. Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4395–4405, 2022. 2

[25] Jongin Lim, Sangdoo Yun, Seulki Park, and Jin Young Choi. Hypergraph-induced semantic tuplet loss for deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 212–222. IEEE, 2022. 6

[26] Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. Chatqa: Surpassing GPT-4 on conversational QA and RAG. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 2

[27] Olga Moskvyak, Frédéric Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Keypoint-aligned embeddings for image retrieval and re-identification. In *Winter Conference on Applications of Computer Vision*, pages 676–685. IEEE, 2021. 2

[28] Bolin Ni, Hongbo Zhao, Chenghao Zhang, Ke Hu, Gaofeng Meng, Zhaoxiang Zhang, and Shiming Xiang. Enhancing visual continual learning with language-guided supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 24068–24077. IEEE, 2024. 2

[29] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 3

[30] Jinhee Park, Jisoo Park, Dagyeong Na, and Junseok Kwon. Deep disentangled metric learning. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 19830–19838. AAAI Press, 2025. 1, 2, 6

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4

[32] Li Ren, Chen Chen, Liqiang Wang, and Kien A. Hua. Learning semantic proxies from visual prompts for parameter-efficient fine-tuning in deep metric learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 1, 6

[33] Karsten Roth, Timo Milbich, Björn Ommer, Joseph Paul Cohen, and Marzyeh Ghassemi. Simultaneous similarity-based self-distillation for deep metric learning. In *Proceedings of Machine Learning Research*, pages 9095–9106. PMLR, 2021. 2

[34] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Non-isotropy regularization for proxy-based deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7410–7420. IEEE, 2022. 2, 6

[35] Mert Bülent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VIII*, pages 153–170. Springer, 2020. 2

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823. IEEE Computer Society, 2015. 5

[37] Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. Learning by planning: Language-guided global image editing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13590–13599. Computer Vision Foundation / IEEE, 2021. 2

[38] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012. IEEE Computer Society, 2016. 5

[39] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[40] Eu Wern Teh, Terrance DeVries, Graham W. Taylor, and Graham. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In *ECCV*, pages 448–464. Springer, 2020. 6

[41] Chengkun Wang, Wenzhao Zheng, Junlong Li, Jie Zhou, and Jiwen Lu. Deep factorized metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 7672–7682. IEEE, 2023. 6

[42] Chengkun Wang, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Introspective deep metric learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(4):1964–1980, 2024. 6

[43] Dongsheng Wang, Miaoge Li, Xinyang Liu, MingSheng Xu, Bo Chen, and Hanwang Zhang. Tuning multi-mode token-level prompt alignment across modalities. *Advances in Neural Information Processing Systems*, 36:52792–52810, 2023. 2

[44] Shijie Wang, Zhihui Wang, Haojie Li, and Wanli Ouyang. Category-specific semantic coherency learning for fine-grained image recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 174–183, 2020. 1

[45] Shijie Wang, Haojie Li, Zhihui Wang, and Wanli Ouyang. Dynamic position-aware network for fine-grained image recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2791–2799, 2021. 1

[46] Shijie Wang, Zhihui Wang, Haojie Li, and Wanli Ouyang. Category-specific nuance exploration network for fine-grained object retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2513–2521, 2022. 2

[47] Shijie Wang, Jianlong Chang, Haojie Li, Zhihui Wang, Wanli Ouyang, and Qi Tian. Learning to parameterize visual attributes for open-set fine-grained retrieval. *Advances in Neural Information Processing Systems*, 36:64681–64694, 2023. 2

[48] Shijie Wang, Jianlong Chang, Haojie Li, Zhihui Wang, Wanli Ouyang, and Qi Tian. Open-set fine-grained retrieval via prompting vision-language evaluator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19381–19391, 2023. 2

[49] Shijie Wang, Jianlong Chang, Zhihui Wang, Haojie Li, Wanli Ouyang, and Qi Tian. Fine-grained retrieval prompt tuning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2644–2652, 2023. 2

[50] Shijie Wang, Jianlong Chang, Zhihui Wang, Haojie Li, Wanli Ouyang, and Qi Tian. Content-aware rectified activation for zero-shot fine-grained image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4366–4380, 2024. 2

[51] Shijie Wang, Zhihui Wang, Haojie Li, Jianlong Chang, Wanli Ouyang, and Qi Tian. Accurate fine-grained object recognition with structure-driven relation graph networks. *International Journal of Computer Vision*, 132(1):137–160, 2024. 1

[52] Shijie Wang, Jian Shi, and Haojie Li. Adversarial reconstruction feedback for robust fine-grained generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3080–3090, 2025. 2

[53] Zhuhui Wang, Shijie Wang, Haojie Li, Zhi Dou, and Jianjun Li. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12289–12296, 2020. 1

[54] Zhihui Wang, Shijie Wang, Shuhui Yang, Haojie Li, Jianjun Li, and Zezhou Li. Weakly supervised fine-grained image classification via guassian mixture model oriented discriminative learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9749–9758, 2020. 1

[55] Zijian Wang, Zheng Zhang, Yandan Luo, Zi Huang, and Heng Tao Shen. Deep collaborative discrete hashing with semantic-invariant structure construction. *IEEE transactions on multimedia*, 23:1274–1286, 2020. 2

[56] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 834–843, 2021. 1

[57] Zijian Wang, Yadan Luo, Zi Huang, and Mahsa Baktashmotlagh. Ffm: Injecting out-of-domain knowledge via factorized frequency modification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4135–4144, 2023. 1

[58] Xiu-Shen Wei, Yang Shen, Xuhao Sun, Han-Jia Ye, and Jian Yang. A^2-net: Learning attribute-aware hash codes for large-scale fine-grained image retrieval. *Advances in Neural Information Processing Systems*, 34:5720–5730, 2021. 1, 2

[59] Yunhao Xu, Zhentao Chen, and Junlin Hu. Deep metric learning in projected-hypersphere space. *Pattern Recognit.*, 161:111245, 2025. 6

[60] Bailin Yang, Haoqiang Sun, Frederick W. B. Li, Zheng Chen, Jianlu Cai, and Chao Song. HSE: hybrid species embedding for deep metric learning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11013–11023. IEEE, 2023. 6

[61] 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. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2

[62] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021. 2

[63] Wenliang Zhao, Yongming Rao, Jie Zhou, and Jiwen Lu. DIML: deep interpretable metric learning via structural matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(4): 2518–2532, 2024. 6

[64] Wenzhao Zheng, Chengkun Wang, Jiwen Lu, and Jie Zhou. Deep compositional metric learning. In *CVPR*, pages 9320–9329. Computer Vision Foundation / IEEE, 2021. 2

[65] Wenzhao Zheng, Borui Zhang, Jiwen Lu, and Jie Zhou. Deep relational metric learning. In *ICCV*, pages 12045–12054. IEEE, 2021.

[66] Xiawu Zheng, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, Feiyue Huang, and Yanhua Yang. Centralized ranking loss with weakly supervised localization for fine-grained object retrieval. In *IJCAI*, pages 1226–1233. ijcai.org, 2018. 2

[67] Haofeng Zhong, Yuchen Hong, Shuchen Weng, Jinxiu Liang, and Boxin Shi. Language-guided image reflection separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 24913–24922. IEEE, 2024. 2

[68] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 4