TRANSGCD: FINE-GRAINED SEMANTIC TRANSFER FOR GENERALIZED CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

Generalized Category Discovery (GCD) aims to uncover novel categories from unlabeled data while preserving recognition of known categories. Recent GCD methods have achieved substantial performance gains by generating pseudo-text representations for unlabeled samples to provide additional semantic guidance. However, these methods typically derive pseudo-text from global visual representations, which are susceptible to background interference and fail to capture fine-grained object attributes. To address these limitations, we propose **TransGCD**, a fine-grained semantic transfer framework that transfers object-centric visual semantics to textual representations for enhanced pseudo-text generation. Specifically, we introduce two complementary techniques, termed **O**bject-centric **S**emantic **D**ecomposition **(OSD)** and **A**ttribute-guided **S**emantic **I**njection **(ASI)**. OSD decomposes global visual representations into complementary object-centric semantics, isolating category-relevant cues from background interference. Building on these decomposed semantics, ASI transfers fine-grained visual attributes into textual cues and injects them into the text encoder to enhance pseudo-text representations. By transferring fine-grained visual semantics into the textual space, TransGCD provides more discriminative semantic guidance for recognizing known categories and discovering novel ones. Extensive experiments across multiple GCD benchmarks demonstrate consistent improvements in known, novel, and overall accuracy, with particularly notable gains on fine-grained datasets. e.g., achieving increases of 8.7% on CUB and 6.4% on Stanford Cars. Code is available at https://anonymous.4open.science/r/TransGCD.

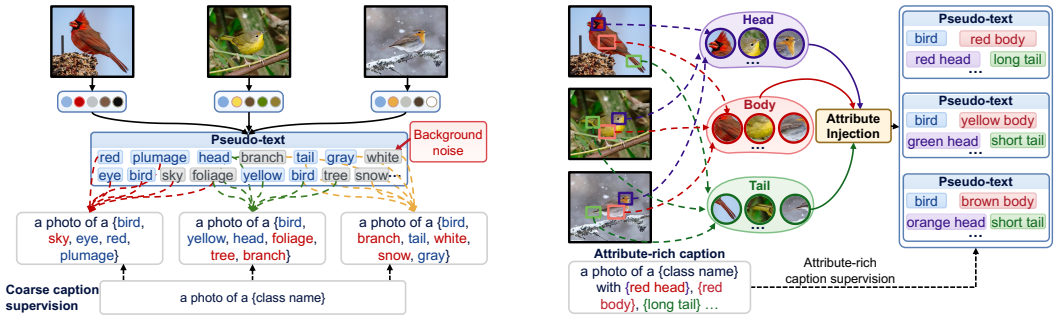
1 INTRODUCTION

Generalized Category Discovery (GCD) (Vaze et al., 2022) is an open-world learning problem that aims to recognize known categories and discover novel ones from unlabeled data. Unlike traditional semi-supervised learning, which assumes labeled and unlabeled data share a predefined category space, GCD allows novel categories to emerge in the unlabeled set. Specifically, labeled samples are available only for known categories, whereas the unlabeled set contains an unknown mixture of known- and novel-category samples. The model must leverage supervision from known categories to recognize known samples, while organizing novel-category samples into coherent clusters, making GCD a realistic yet challenging setting for visual recognition.

State-of-the-art GCD methods increasingly construct instance-specific pseudo-text from visual representations to provide additional semantic supervision for unlabeled data (Wang et al., 2025; Han et al., 2026). Specifically, GET (Wang et al., 2025) maps the global image representation into pseudo tokens and processes them with the CLIP text encoder. Meanwhile, AL-GCD (Han et al., 2026) uses global visual similarities to aggregate label-text embeddings from relevant labeled samples. By transferring visual semantics into the textual space, these methods have achieved strong performance in category discovery. However, their reliance on global visual representations introduces two critical limitations. First, the resulting pseudo-text may contain irrelevant background information. As illustrated in Fig. 1a, scene concepts such as *sky*, *branches*, and *snow* are propagated into the pseudo-text despite providing little evidence for identifying the object category. Second, encoding the entire image as a holistic representation limits the explicit modeling of part-level attributes essential for distinguishing visually similar categories.

To address these issues, we propose **TransGCD**, a fine-grained semantic transfer framework that enhances pseudo-text generation by transferring object-centric visual semantics into textual representations. TransGCD consists of two complementary modules: **O**bject-centric **S**emantic **D**ecomposition (**OSD**) and **A**ttribute-guided **S**emantic **I**njection (**ASI**). Specifically, OSD employs multiple learnable semantic queries to decompose global visual representations into complementary object-centric semantics. However, interactions with the entire image may cause these queries to aggregate irrel-

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.



(a) Existing Methods (b) Ours TransGCD

Figure 1: Conceptual comparison of pseudo-text generation. (a) Existing methods (Wang et al., 2025; Han et al., 2026) rely on holistic visual features, introducing background noise and overlooking part-level attributes. (b) TransGCD transfers object-centric attributes into the textual space, producing cleaner and more discriminative pseudo-text.

evant background information. To mitigate this issue, OSD introduces a saliency-guided mask that suppresses irrelevant patches and guides the queries to attend category-relevant image regions. As a result, OSD captures part-level cues of objects with reduced interference from background. Building on these decomposed visual semantics, ASI injects fine-grained attributes into the text encoder to produce more discriminative pseudo-text representations. As illustrated in Fig. 1b, the refined pseudo-text captures discriminative attributes of different object parts, such as the *head*, *body*, and *tail*, rather than irrelevant background concepts. Experiments on six benchmarks show that TransGCD consistently improves existing GCD methods, with an average gain of **4.4%** overall and **7.0%** on fine-grained benchmarks.

Our contributions are summarized as follows: (1) We propose **TransGCD**, a fine-grained semantic transfer framework that enhances pseudo-text representations with visual semantics for GCD. (2) We propose **Object-centric Semantic Decomposition (OSD)**, which extracts fine-grained attributes from visual representations while suppressing background interference. (3) We propose **Attribute Semantic Injection (ASI)**, which transfers extracted attributes into the textual space to generate richer and more discriminative pseudo-text representations. (4) We conduct extensive experiments on six GCD benchmarks with multiple baselines. TransGCD consistently improves overall, known-, and novel-class accuracy, with notable gains on fine-grained datasets.

2 Related Work

2.1 Generalized Category Discovery

Generalized Category Discovery (GCD) (Vaze et al., 2022) aims to recognize known categories and discover novel ones from unlabeled data. Most methods build upon DINO representations: SimGCD (Wen et al., 2023) establishes a strong parametric baseline, CMS (Choi et al., 2024) integrates mean-shift clustering with contrastive learning, and SelEx (Rastegar et al., 2024) introduces self-expertise learning with hierarchical pseudo-labeling. While these methods improve visual representations, recent studies incorporate textual semantics as complementary supervision. TextGCD (Zheng et al., 2024), SSR^2-GCD (He et al., 2026), and SpectralGCD (Caselli et al., 2026) construct pseudo-text by filtering concepts from predefined lexicons. Moving beyond retrieval, GET (Wang et al., 2025) directly maps global image representations into pseudo tokens, whereas AL-GCD (Han et al., 2026) transfers class-level textual semantics to unlabeled samples according to global visual similarities. Although these methods demonstrate the effectiveness of pseudo-text supervision, their reliance on holistic visual representations may propagate background noise into the textual space and overlook discriminative part-level attributes.

2.2 Vision-Language Pre-trained Models

Vision–language models (VLMs) learn transferable multimodal representations from large-scale image–text pairs. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) employ contrastive

dual encoders to align global visual and textual representations in a shared semantic space. Subsequent methods further improve cross-modal learning: ALBEF (Li et al., 2021) combines unimodal alignment with cross-modal fusion, while BLIP (Li et al., 2022) improves pretraining by bootstrapping captions from noisy web data. Among these models, CLIP is particularly influential due to its strong zero-shot transfer and open-vocabulary recognition, and can be efficiently adapted through prompt learning (Zhou et al., 2022b;a). However, conventional CLIP adaptation relies on holistic image representations and coarse class-level prompts, limiting supervision for fine-grained object attributes. Recent studies therefore explore finer-grained image–text alignment. FLAIR (Xiao et al., 2025) learns localized visual representations from detailed captions, whereas FineLIP (Asokan et al., 2025) aligns local visual tokens with fine-grained textual descriptions. Motivated by these advances, TransGCD enriches class-level supervision with attribute-rich descriptions of complementary object parts, enabling more discriminative pseudo-text generation.

3 METHOD

3.1 PRELIMINARIES

Problem formulation. Generalized Category Discovery (GCD) aims to discover novel categories from unlabeled data while maintaining recognition of known categories. Formally, we are given a labeled dataset $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l} \subset \mathcal{X} \times \mathcal{Y}^l$, where $\mathcal{Y}^l$ denotes the set of known categories, and an unlabeled dataset $\mathcal{D}^u = \{x_i^u\}_{i=1}^{N_u} \subset \mathcal{X}$, which contains samples from both known and novel categories. Let $\mathcal{Y}^u$ denote the categories in $\mathcal{D}^u$, satisfying $\mathcal{Y}^l \subset \mathcal{Y}^u$. The goal is to assign each sample in $\mathcal{D}^u$ to its corresponding known or novel category.

Overview. TransGCD generates discriminative pseudo-text by transferring fine-grained visual semantics into the textual space. As illustrated in Fig. 2, it comprises two stages: *fine-grained pseudo-text learning* and *plug-and-play GCD training*. Specifically, Stage I learns a pseudo-text generator from $\mathcal{D}^l$ using two complementary modules. OSD (Sec. 3.2) decomposes global representations into object-centric semantics, while ASI (Sec. 3.3) transfers them into fine-grained pseudo-text. Subsequently, Stage II applies the generator to both $\mathcal{D}^l$ and $\mathcal{D}^u$ and fuses its outputs with global visual features for plug-and-play GCD training.

3.2 OBJECT-CENTRIC SEMANTIC DECOMPOSITION

As illustrated in Fig. 1(a), existing pseudo-text-based GCD methods derive textual semantics from holistic visual representations. As a result, the resulting pseudo-text may capture category-relevant object semantics (e.g., a bird's head and wings) while also encoding irrelevant background noise (e.g., sky and branches). To address this issue, we propose **Object-Centric Semantic Decomposition (OSD)**. Specifically, OSD employs learnable part queries to decompose global visual representations into complementary object-centric semantics. As a result, OSD yields cleaner object-centric visual priors for subsequent pseudo-text generation.

Query-driven attribute discovery. To capture diverse fine-grained attributes from an image, we introduce P learnable queries $Q = [q_1, \ldots, q_P] \in \mathbb{R}^{P \times D}$, with each query learning to capture the visual representation of a distinct attribute (e.g., head, body, tail, etc). Given an image x_i and its corresponding patch tokens $V_i = [v_{i,1}, \ldots, v_{i,N}] \in \mathbb{R}^{N \times D}$ from CLIP visual encoder. We then concatenate the queries with the patch tokens to form $Z_i^0 = [Q; V_i] \in \mathbb{R}^{(P+N) \times D}$, and feed them into a Transformer encoder:

$$Z_i^l = \mathcal{F}_l(Z_i^{l-1}), \quad l = 1, \ldots, L_o - 1, \tag{1}$$

where $\mathcal{F}_l$ is the l-th block and L_o is the total number of OSD Transformer blocks. Through self-attention, the queries aggregate local visual semantics from the patch tokens. However, they may capture irrelevant background cues. Therefore, we apply a saliency-guided mask in the final block to further suppress background noise.

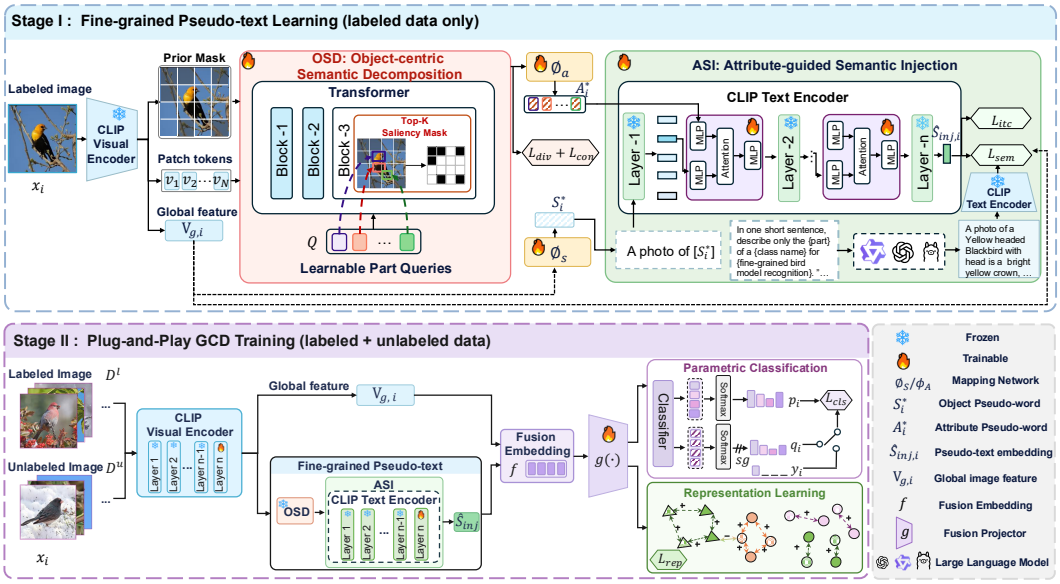


Figure 2: Overview of our proposed TransGCD framework, which refines pseudo-text representations with object-centric attributes for GCD. (a) Stage 1 learns a fine-grained pseudo-text generator, where OSD extracts object-centric attributes A_i^* and ASI injects them into the object pseudo-word S_i^* to produce $S_{inj,i}^*$. (b) Stage 2 applies the learned generator to labeled and unlabeled images and fuses $S_{inj,i}^*$ with the global visual feature V_g, i for GCD training.

Saliency-Guided Filtering. We construct a mask using the CLS-to-patch attention of the CLIP visual encoder. Specifically, we average the attention responses across L_c blocks and H heads:

$$\rho_{i,j} = \frac{1}{L_c H} \sum_{\ell=1}^{L_c} \sum_{h=1}^{H} a_{i,j}^{\ell,h}, \qquad j = 1, \ldots, N, \tag{2}$$

where $a_{i,j}^{\ell,h}$ is the CLS-to-patch attention and $\rho_{i,j}$ is the corresponding saliency score. We then select the top-K salient patches to construct the mask:

$$m_{i,j} = \begin{cases} 0, & j \in \text{TopK}\left(\{\rho_{i,n}\}_{n=1}^{N}, K\right), \\ -\infty, & \text{otherwise.} \end{cases} \tag{3}$$

The mask m_i is added to the part-query-to-patch attention logits in the final OSD block. The resulting object-centric attributes are

$$A_i = \left[\mathcal{F}_{L_o}\left(Z_i^{L_o-1}; m_i\right)\right]_{[1:P,:]} \in \mathbb{R}^{P \times D}. \tag{4}$$

Part-Wise Semantic Regularization. To encourage the attributes A_i to capture part-wise semantics across different objects, we introduce two complementary losses. The cross-image consistency loss $\mathcal{L}_{\text{con}}$ encourages the same query to produce similar attribute representations for images of the same class. The within-image diversity loss $\mathcal{L}_{\text{div}}$ encourages different queries in each image to produce distinct representations. Let $z_{i,p} = A_{i,p}/\|A_{i,p}\|_2$ denote the normalized representation of the p-th query for image x_i. We first enforce cross-image consistency with a supervised contrastive loss:

$$\mathcal{L}_{\text{con}} = -\frac{1}{PB} \sum_{p=1}^{P} \sum_{i=1}^{B} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp(z_{i,p}^{\top} z_{j,p}/\tau_s)}{\sum_{k \neq i} \exp(z_{i,p}^{\top} z_{k,p}/\tau_s)}, \tag{5}$$

where $\mathcal{P}(i) = \{j \neq i \mid y_j = y_i\}$. We then encourage within-image diversity by minimizing the similarity between different queries:

$$\mathcal{L}_{\text{div}} = \frac{1}{BP(P-1)} \sum_{i=1}^{B} \sum_{p=1}^{P} \sum_{\substack{q=1 \\ q \neq p}}^{P} \left(z_{i,p}^{\top} z_{i,q}\right)^2. \tag{6}$$

3.3 ATTRIBUTE-GUIDED SEMANTIC INJECTION

As described in Sec. 3.2, OSD decomposes the global visual representation into complementary fine-grained semantics, such as the head, body, and tail of a bird, which help distinguish visually similar categories. However, they remain in the visual space and cannot directly refine pseudo-text representations. Therefore, we introduce **Attribute-guided Semantic Injection (ASI)** to transfer these attributes into the CLIP text space and integrate them with global object semantics. As a result, ASI produces more discriminative fine-grained pseudo-text representations.

Attribute Mapping. We introduce a lightweight mapping network ϕ_a to project the decomposed attributes into the text embedding space. Specifically, for the i-th image, given the attribute semantics $A_i \in \mathbb{R}^{P\times D}$ from OSD, we obtain:

$$A_i^* = \phi_a(A_i) \in \mathbb{R}^{P\times D}. \tag{7}$$

Here, A_i^* is regarded as P attribute pseudo-word tokens which describe the fine-grained visual cues.

Attribute Semantic Injection. The existing pseudo-text-based GCD method (Wang et al., 2025) generates pseudo-text from global visual representations. In contrast, we retain the global semantics as an anchor and introduce fine-grained attribute semantics to enrich the pseudo-text representation. First, we map the global visual representation $V_{g,i}$ into the CLIP text space:

$$S_i^* = \phi_s(V_{g,i}) \in \mathbb{R}^D. \tag{8}$$

Then, we insert S_i^* into the template "a photo of S_i^*" and feed it into the CLIP text encoder, yielding the hidden representation $S_i^{(l)}$ at layer l. Next, we inject the attribute tokens A_i^* into $S_i^{(l)}$. Specifically, $S_i^{(l)}$ and A_i^* are first projected by lightweight bottleneck MLPs $\psi_s^{(l)}$ and $\psi_a^{(l)}$, respectively:

$$\bar{S}_i^{(l)} = \psi_s^{(l)}(S_i^{(l)}), \qquad \bar{A}_i^{(l)} = \psi_a^{(l)}(A_i^*). \tag{9}$$

We then apply multi-head cross-attention, using $\bar{S}_i^{(l)}$ as the query and $\bar{A}_i^{(l)}$ as the keys and values, to aggregate attribute information:

$$C_i^{(l)} = \mathrm{MHA}\left(\bar{S}_i^{(l)}, \bar{A}_i^{(l)}, \bar{A}_i^{(l)}\right), \tag{10}$$

where $C_i^{(l)}$ denotes the attribute representation that aggregates the semantics of relevant local regions. The resulting attribute information is projected by an output MLP $\psi_u^{(l)}$ and injected into $S_i^{(l)}$ through a residual connection:

$$\hat{S}_{inj,i}^{(l)} = S_i^{(l)} + \psi_u^{(l)}\left(C_i^{(l)}\right), \tag{11}$$

where $\hat{S}_{inj,i}^{(l)}$ denotes the updated representation at the l-th injection layer. After successive injections, the final pseudo-text $\hat{S}_{inj,i}$ integrates global object semantics with local attributes.

Image-Text Alignment Loss. To further preserve image-specific semantics, we align $\hat{S}_{inj,i}$ with the corresponding global visual representation $V_{g,i}$ using a symmetric contrastive objective. For a batch of B images, let $s_{ij} = \mathrm{sim}\left(V_{g,i}, \hat{S}_{inj,j}\right)/\tau$. The image-pseudo-text contrastive loss is:

$$\mathcal{L}_{\mathrm{itc}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ij})} + \log\frac{\exp(s_{ii})}{\sum_{j=1}^{B}\exp(s_{ji})}\right], \tag{12}$$

where τ is the temperature. This objective aligns each image with its corresponding pseudo-text while separating mismatched pairs.

3.4 LLM-GUIDED FINE-GRAINED SUPERVISION

To enrich category-level supervision with fine-grained semantics, we introduce an LLM-guided target construction pipeline. As illustrated in Fig. 3, it mines part attributes, grounds them with category visual prototypes, and constructs attribute-rich semantic targets.

Attribute Mining and Grounding. For each category c, we first prompt an LLM to propose a set of object parts $\mathcal{P}_c = \{p_1, \ldots, p_{M_c}\}$, where $M_c = |\mathcal{P}_c|$. For each $p \in \mathcal{P}_c$, the LLM describes its visual appearance, and spaCy (Honnibal et al., 2020) extracts a set of candidate attributes $\mathcal{W}_{c,p}$. To retain visually representative attributes, we compute the category prototype and perform visual grounding as

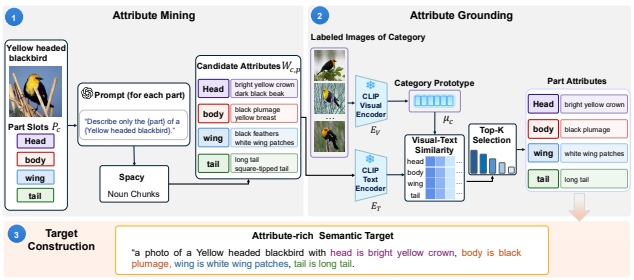


Figure 3: Overview of LLM-guided target construction.

$$\mu_c = \mathrm{Norm}\left(\frac{1}{|\mathcal{D}_c^l|}\sum_{x_i \in \mathcal{D}_c^l} E_V(x_i)\right), \qquad \mathcal{W}_{c,p}^* = \mathrm{TopK}_{w \in \mathcal{W}_{c,p}} \mathrm{sim}\left(E_T(w), \mu_c\right), \tag{13}$$

where $\mathcal{D}_c^l$ is the labeled image set of category c, and E_V and E_T are the frozen CLIP image and text encoders. Thus, $\mathcal{W}_{c,p}^*$ contains the visually grounded attributes of part p.

Semantic Target Construction. Next, we combine the category name c with the grounded part–attribute pairs $\{(p, \mathcal{W}_{c,p}^*)\}_{p \in \mathcal{P}_c}$ to construct a target $\mathcal{T}_c$. The target follows the template "a photo of a c with p_1 is $\mathrm{Join}(\mathcal{W}_{c,p_1}^*)$, ..., and p_{M_c} is $\mathrm{Join}(\mathcal{W}_{c,p_{M_c}}^*)$," where $\mathrm{Join}(\cdot)$ concatenates the selected attributes into a phrase. We encode the target as $t_c = E_T(\mathcal{T}_c)$ and supervise $\hat{S}_{inj,i}$ by

$$\mathcal{L}_{\mathrm{sem}} = \frac{1}{B}\sum_{i=1}^{B}\left[1 - \mathrm{sim}\left(\hat{S}_{inj,i}, t_{y_i}\right)\right]. \tag{14}$$

Training Objective. The overall objective for fine-grained pseudo-text learning integrates the above regularization and alignment objectives:

$$\mathcal{L}_{\mathrm{Stage1}} = \lambda\left(\mathcal{L}_{\mathrm{con}} + \mathcal{L}_{\mathrm{div}} + \mathcal{L}_{\mathrm{itc}}\right) + \beta\mathcal{L}_{\mathrm{sem}}, \tag{15}$$

where λ and β control the contributions of the corresponding objectives.

3.5 Plug-and-Play GCD Training

Feature Fusion. After Stage I, we apply the learned pseudo-text generator to all samples in $\mathcal{D}^l \cup \mathcal{D}^u$, obtaining pseudo-text representation $\hat{S}_{inj,i}$ for each image x_i. We fuse them as

$$f_i = g\left(\alpha V_{g,i} + (1-\alpha)\hat{S}_{inj,i}\right), \tag{16}$$

where $\alpha = 0.5$ follows Caselli et al. (2026), $g(\cdot)$ is the projection head, and f_i is the fused feature.

GCD training. Our fused representation f_i directly replaces the visual features in existing GCD frameworks, providing fine-grained textual semantics without altering their training objectives. Taking SimGCD (Wen et al., 2023) as an example in Fig. 2, we optimize

$$\mathcal{L}_{\mathrm{Stage2}} = \mathcal{L}_{\mathrm{rep}} + \mathcal{L}_{\mathrm{cls}}, \tag{17}$$

where $\mathcal{L}_{\mathrm{rep}}$ and $\mathcal{L}_{\mathrm{cls}}$ denote the representation and classification losses, respectively.

4 Experiment

4.1 Experimental Setup

Datasets. We evaluate TransGCD on six widely used benchmarks, including two generic datasets, CIFAR-100 (Krizhevsky et al., 2009) and ImageNet-100 (Deng et al., 2009), and four fine-grained datasets, CUB-200 (Wah et al., 2011), Stanford Cars (Krause et al., 2013), FGVC-Aircraft (Maji et al., 2013), and Herbarium19 (Tan et al., 2019). Herbarium19 further provides a challenging setting with long-tailed fine-grained categories with large per-category sample imbalance. Following the standard GCD protocol (Vaze et al., 2022), 50% of the seen categories are randomly selected as labeled data $\mathcal{D}^l$, while the remaining seen and novel samples form the unlabeled set $\mathcal{D}^u$.

Evaluation Protocols and Implementation Details. Following standard GCD protocols (Vaze et al., 2022; Wen et al., 2023), we use clustering accuracy (ACC). Let y_i and $\hat{y}_i$ denote the ground-truth label and predicted cluster assignment of the i-th sample. ACC is computed as $\text{ACC} = \frac{1}{|\mathcal{D}^u|}\sum_{i=1}^{|\mathcal{D}^u|}\mathbb{I}[y_i = \pi^\star(\hat{y}_i)]$, where $\mathbb{I}[\cdot]$ is the indicator function and $\pi^\star$ is the optimal assignment obtained by the Hungarian algorithm (Kuhn, 1955). We follow Wang et al. (2025); Caselli et al. (2026) and use CLIP ViT-B/16 as the backbone. The learning rate, batch size, temperature τ_s, and weight decay are set to 0.1, 128, 0.07, and 1×10^{-4}, respectively. We train for 200 epochs, with 3 Transformer layers in OSD and 2 cross-attention layers in ASI. We select the top-50% salient patches for the mask and the top-3 noun phrases extracted by spaCy for the target prompt. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU.

4.2 Comparison with State-of-the-Art Methods

We compare TransGCD with DINO-based visual-only methods, including SimGCD (Wen et al., 2023), CMS (Choi et al., 2024), and SelEx (Rastegar et al., 2024). For CLIP-based methods, we integrate TransGCD into SimGCD-CLIP, CMS-CLIP, and SelEx-CLIP. We report overall, known, and novel-class accuracy across six benchmarks.

Evaluation on fine-grained datasets. As shown in Table 1, TransGCD consistently improves all three baselines across CUB, Stanford Cars, and FGVC-Aircraft. Notably, integrating TransGCD into SelEx-CLIP raises overall accuracy from 68.5% to 87.9% on CUB and from 63.1% to 69.9% on FGVC-Aircraft. Averaged across the baselines, TransGCD improves novel-class accuracy by **12.1**, **8.2**, and **5.9** points on the three datasets, respectively. Across all baseline–dataset combinations, it achieves average gains of **7.0**, **3.6**, and **8.7** points in overall, known-class, and novel-class accuracy.

Table 1: Evaluation on the Semantic Shift Benchmark (SSB). Bold values represent the best results, while underlined values represent the second-best results.

	Method	CUB			Stanford Cars			FGVC-Aircraft		
		All	Old	New	All	Old	New	All	Old	New
DINO	GCD (Vaze et al., 2022)	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9
	SimGCD (Wen et al., 2023)	60.3	65.6	57.7	53.8	71.9	45.0	54.2	59.1	51.8
	CMS (Choi et al., 2024)	68.2	76.5	64.0	56.9	76.1	47.6	56.0	63.4	52.3
	LegoGCD (Cao et al., 2024)	63.8	71.9	59.8	57.3	75.7	48.4	55.0	61.5	51.7
	SPTNet (Wang et al., 2024)	65.8	68.8	65.1	59.0	79.2	49.3	59.3	61.8	58.1
	SelEx (Rastegar et al., 2024)	73.6	75.3	72.8	58.5	75.6	50.3	57.1	64.7	53.3
	AllGCD (Cao et al., 2025)	68.4	75.1	65.1	60.5	77.0	52.5	57.4	62.4	54.9
	AF (Xu et al., 2025)	69.0	74.3	66.3	67.0	80.7	60.4	59.4	68.1	55.0
	AptGCD (Zhang et al., 2025)	70.3	74.3	69.2	62.1	79.7	53.6	61.1	65.2	59.0
	APL (Dai et al., 2025)	64.5	68.1	62.1	60.1	77.6	51.2	56.6	60.2	54.8
	ConGCD (Tang et al., 2025)	81.7	80.4	82.4	62.5	70.2	58.7	57.5	**77.5**	47.9
	SEAL (He et al., 2025)	66.2	72.1	63.2	65.3	79.3	58.5	62.0	65.3	60.4
	EAGC (Zheng et al., 2026)	83.2	73.9	87.9	65.7	83.7	57.0	65.7	65.6	65.7
	BIA (Tang et al., 2026)	80.6	**81.0**	80.4	57.0	77.3	47.2	61.8	68.2	59.2
CLIP	CPT (Yang et al., 2025b)	70.1	73.5	68.4	74.2	84.3	69.3	53.3	56.1	51.9
	GET (Wang et al., 2025)	77.0	78.1	76.4	78.5	86.8	74.5	58.9	59.6	58.5
	SSR^2-GCD (He et al., 2026)	78.3	78.5	78.2	**89.2**	**93.1**	87.3	–	–	–
	SpectralGCD (Caselli et al., 2026)	79.2	80.4	78.5	89.1	92.6	**87.4**	63.0	66.1	61.4
	AL-GCD (Han et al., 2026)	84.1	79.7	86.3	79.0	90.8	73.6	66.6	66.1	66.8
	SimGCD-CLIP*	70.3	75.8	67.5	68.3	84.0	60.8	53.2	53.5	53.0
	+ TransGCD (Ours)	73.7	77.5	71.8	76.5	84.9	72.5	57.2	61.4	55.0
	CMS-CLIP*	65.8	75.9	60.7	75.0	87.7	68.9	48.1	52.2	46.0
	+ TransGCD (Ours)	69.1	75.4	66.8	76.4	90.0	69.9	54.8	60.6	51.9
	SelEx-CLIP*	68.5	72.5	66.5	72.6	88.3	65.1	63.1	67.0	61.1
	+ TransGCD (Ours)	**87.9**	79.1	**92.3**	82.2	92.8	77.1	**69.9**	67.9	**70.9**
	Avg. Δ	**+8.7**	**+2.6**	**+12.1**	**+6.4**	**+2.6**	**+8.2**	**+5.8**	**+5.7**	**+5.9**

* denotes results reproduced by us.

Evaluation on generic and long-tailed datasets. As shown in Table 2, TransGCD consistently improves the overall accuracy of all baselines on CIFAR-100 and ImageNet-100. When integrated into SelEx-CLIP, it achieves 81.7% and 93.0% overall accuracy, with novel-class gains of **4.1** and **4.0** points, respectively. Likewise, SimGCD-CLIP+TransGCD improves novel-class accuracy by **12.8** and **2.2** points. On the long-tailed Herbarium19 dataset, TransGCD yields average gains of **1.3** and **2.8** points in overall and novel-class accuracy. Across all three datasets, CMS-CLIP+TransGCD improves these two metrics by an average of **2.3** and **6.2** points, respectively.

Table 2: Evaluation on two generic image recognition datasets and Herbarium19. Bold values represent the best results, while underlined values represent the second-best results.

	Method	CIFAR-100			ImageNet-100			Herbarium19		
		All	Old	New	All	Old	New	All	Old	New
DINO	GCD (Vaze et al., 2022)	73.0	76.2	66.5	74.1	89.8	66.3	35.4	51.0	27.0
	SimGCD (Wen et al., 2023)	80.1	81.2	77.8	83.0	93.1	77.9	44.0	58.0	36.4
	CMS (Choi et al., 2024)	82.3	85.7	75.5	84.7	95.6	79.2	36.4	54.9	26.4
	LegoGCD (Cao et al., 2024)	81.8	81.4	82.5	86.3	94.5	82.1	45.1	57.4	38.4
	SPTNet (Wang et al., 2024)	81.3	84.3	75.6	85.4	93.2	81.4	43.4	58.7	35.2
	SelEx (Rastegar et al., 2024)	82.3	84.2	81.9	86.9	94.2	83.2	46.4	61.2	38.4
	AllGCD (Cao et al., 2025)	82.3	82.3	82.2	86.5	94.7	82.3	45.2	58.3	38.0
	AF (Xu et al., 2025)	82.2	85.0	76.5	85.4	94.6	80.8	45.5	59.0	38.3
	AptGCD (Zhang et al., 2025)	82.8	81.8	85.5	87.8	95.4	84.3	48.4	60.6	42.3
	APL (Dai et al., 2025)	80.9	–	79.5	83.2	–	78.5	44.9	58.1	37.9
	ConGCD (Tang et al., 2025)	80.5	84.9	71.7	83.5	94.2	78.0	–	–	–
	SEAL (He et al., 2025)	82.1	81.7	83.0	84.6	90.9	81.3	46.9	45.8	**48.2**
	EAGC (Zheng et al., 2026)	79.3	84.1	69.7	89.8	96.2	86.5	42.0	55.0	35.0
	BIA (Tang et al., 2026)	80.7	84.3	72.1	82.8	94.1	77.8	36.8	47.5	31.0
CLIP	CPT (Yang et al., 2025b)	81.3	81.3	81.2	89.2	95.2	86.1	45.3	61.2	36.7
	GET (Wang et al., 2025)	82.1	85.5	75.5	91.7	95.7	89.7	**49.7**	64.5	41.7
	SSR^2-GCD (He et al., 2026)	**86.4**	86.2	**86.9**	92.1	96.0	90.2	–	–	–
	SpectralGCD (Caselli et al., 2026)	86.1	**87.2**	83.9	**93.4**	96.1	**92.0**	–	–	–
	AL-GCD (Han et al., 2026)	80.4	85.6	69.8	92.7	96.2	91.1	47.7	59.0	41.6
	SimGCD-CLIP*	78.4	84.0	67.2	90.5	95.2	88.2	47.4	64.3	38.3
	+ TransGCD (Ours)	81.8	82.7	80.0	91.7	94.3	90.4	48.3	**65.3**	39.2
	CMS-CLIP*	80.8	86.3	69.7	89.4	96.5	85.8	36.9	57.5	25.8
	+ TransGCD (Ours)	83.6	84.2	82.3	89.8	96.3	86.5	40.6	58.5	31.0
	SelEx-CLIP*	80.8	86.0	70.4	89.7	95.2	87.0	45.8	57.0	37.6
	+ TransGCD (Ours)	81.7	85.3	74.5	93.0	**97.1**	91.0	45.2	55.2	39.8
	Avg. Δ	**+2.4**	-1.4	**+9.8**	**+1.6**	**+0.3**	**+2.3**	**+1.3**	**+0.1**	**+2.8**

* denotes results reproduced by us.

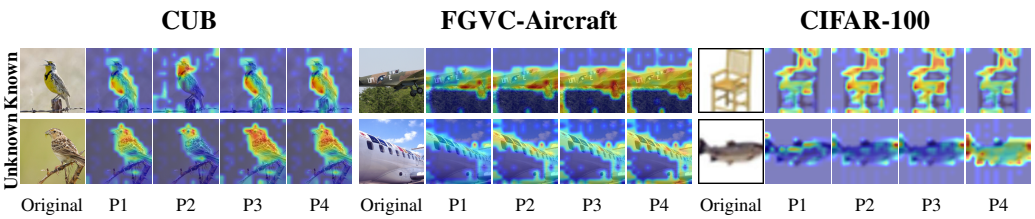


Figure 4: Visualization of part attention maps on fine-grained and generic datasets.

Object-centric semantic visualization. We visualize query-to-patch attention maps to examine the semantics learned by OSD. As shown in Fig. 4, the queries exhibit specialization across known and novel samples. They localize complementary bird parts on CUB, structural components on FGVC-Aircraft, and foreground regions on low-resolution CIFAR-100 images. These patterns indicate that OSD learns transferable object-centric decomposition rather than dataset-specific background correlations, providing cleaner visual priors for pseudo-text construction.

4.3 Ablation Study

Effect of different components. We construct a baseline by mapping the global image representation to a pseudo-word S_i^*, encoding "a photo of S_i^*" and supervising it with "a photo of a {class name}." Building on this baseline, we progressively add the attribute-rich Prompt, OSD, and ASI. Specifically, replacing the coarse target with the part-aware Prompt improves overall accuracy by **0.5** and **0.8** points on CUB and FGVC-Aircraft, respectively. With this supervision, OSD further adds **0.8** and **0.6** points. Subsequently, ASI contributes another **1.2** and **2.1** points. Consequently, the complete model outperforms the baseline by **0.7** and **0.4** points on CIFAR-100 and ImageNet-100. Overall, these consistent gains validate the complementary design and effectiveness of TransGCD.

Effect of mask guidance in OSD. We isolate the effect of mask guidance on CUB and CIFAR-100 while keeping the decomposition architecture fixed. As shown in Table 4, the mask consistently improves category discovery, confirming its effectiveness in suppressing background interference and learning more discriminative object-centric semantics.

Table 4: Effect of mask in OSD.

Dataset	Mask	All	Old	New
CUB	✓	86.7	80.9	89.7
	✗	86.2	81.9	88.4
CIFAR-100	✓	81.3	85.2	73.7
	✗	80.6	84.6	72.5

Table 3: Ablation study of different components in our proposed TransGCD.

Baseline	Prompt	OSD	ASI	CUB			FGVC-Aircraft			CIFAR100			ImageNet-100		
				All	Old	New	All	Old	New	All	Old	New	All	Old	New
✓	✗	✗	✗	85.4	78.4	88.9	66.4	64.2	67.6	81.0	85.6	71.8	92.6	96.8	90.4
✓	✓	✗	✗	85.9	80.0	88.9	67.2	64.8	68.4	81.2	85.2	73.1	92.8	97.1	90.1
✓	✓	✓	✗	86.7	80.9	89.7	67.8	64.2	69.6	81.3	85.2	73.7	92.4	96.9	90.1
✓	✓	✓	✓	87.9	79.1	92.3	69.9	67.9	70.9	81.7	85.3	74.5	93.0	97.1	91.0
Improvement over our baseline				**+ 2.4**	**+0.7**	**+3.4**	**+ 3.6**	**+ 3.7**	**+3.3**	**+ 0.7**	-0.3	**+ 2.7**	**+ 0.4**	**+ 0.3**	**+0.6**

Effect of the injection layer in ASI. We examine ASI at different text-encoder layers. As shown in Table 5, injection at earlier layers generally outperforms deep injection across all four datasets. Among single-layer settings, Layer 3 performs best on CUB and CIFAR-100, while Layer 6 performs best on FGVC-Aircraft and ImageNet-100. This suggests that early injection leaves sufficient depth for progressive attribute integration. Multi-layer injection offers a slight overall gain only on CUB, but provides no consistent advantage on the other datasets.

Table 5: Effect of different ASI injection positions.

Strategy	Layer(s)	CUB			FGVC-Aircraft			CIFAR-100			ImageNet-100		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
Single-layer	3	87.6	83.4	89.8	70.1	66.7	71.8	81.7	85.3	74.5	92.5	96.8	90.4
	6	85.5	79.2	88.6	70.8	64.9	73.7	81.2	85.0	73.6	93.0	97.1	91.0
	9	86.8	81.5	89.1	69.9	67.6	71.0	80.9	84.0	74.8	90.6	96.8	87.5
Multi-layer	3, 6, 9	87.9	79.1	92.3	69.9	67.9	70.9	80.6	84.8	72.2	91.6	96.5	89.1

Effect of textual supervision. We compare directly fused global captions with attribute-rich part descriptions used to supervise TransGCD. The descriptions are generated by two VLMs, Qwen2.5-VL-7B (Bai et al., 2025) and LLaVA-1.5-7B (Liu et al., 2023), and one LLM, Qwen3-8B (Yang et al., 2025a). As shown in Table 6, Qwen2.5-VL part descriptions outperform its global captions by **12.8**, **14.1**, and **18.0** points on CUB, Stanford Cars, and FGVC-Aircraft, respectively. The corresponding gains with LLaVA are **3.4**, **12.9**, and **6.4** points, with an additional **5.3**-point gain on CIFAR-100. Moreover, the text-only Qwen3 remains within **0.5** points of the best VLM result across all datasets. These results demonstrate the effectiveness of part-level supervision and its robustness across generators with different capabilities.

Table 6: Effect of different textual supervision signals.

Text supervision	Generator	CUB			Stanford Cars			FGVC-Aircraft			CIFAR-100		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
Global caption	Qwen2.5-VL-7b (Bai et al., 2025)	74.9	70.4	77.2	67.9	80.0	62.0	52.0	59.8	48.1	81.7	84.7	75.6
	LLaVA-1.5-7B (Liu et al., 2023)	83.4	75.9	87.2	69.4	84.8	61.9	63.1	59.6	64.9	75.9	81.0	65.7
Part description	Qwen2.5-VL-7b (Bai et al., 2025)	87.7	80.7	91.3	82.0	91.6	77.4	70.0	67.7	71.2	80.4	85.5	70.2
	LLaVA-1.5-7b (Liu et al., 2023)	86.8	78.1	91.2	82.3	91.2	78.0	69.5	65.8	71.4	81.2	85.5	72.8
	Qwen3-8b (Yang et al., 2025a)	87.9	79.1	92.3	82.2	92.8	77.1	69.9	67.9	70.9	81.7	85.3	74.5

5 Conclusion

In this work, we identify global visual representations as a key bottleneck in pseudo-text-based GCD, since they introduce background noise and overlook discriminative object attributes. To address this limitation, we propose TransGCD, a fine-grained semantic transfer framework for object-centric pseudo-text generation. Specifically, OSD decomposes global visual representations into complementary object-centric semantics with reduced background interference, while ASI transfers these semantics into the textual space to enrich pseudo-text representations. Extensive experiments on six benchmarks and multiple GCD baselines demonstrate consistent improvements in overall, known-class, and novel-class accuracy, particularly on fine-grained datasets. These results highlight the importance of transferring fine-grained visual semantics into the textual space for more discriminative category discovery.

AI USE STATEMENT

In this work, we used large language models to polish the language and improve the readability of the manuscript. Additionally, we used generative AI tools to generate class-specific part descriptions for constructing textual supervision. We have reviewed all AI-assisted work. AI-assisted manuscript edits were checked by the authors for technical accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Mothilal Asokan, Kebin Wu, and Fatima Albreiki. FineLIP: Extending CLIP's reach via fine-grained alignment with longer text inputs. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14495–14504, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Xinzi Cao, Yutong Lu, and Yonghong Tian. Solving the catastrophic forgetting problem in generalized category discovery. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16880–16889, 2024.

Xinzi Cao, Ke Chen, Feidiao Yang, Xiawu Zheng, Yonghong Tian, and Yutong Lu. AllGCD: Leveraging all unlabeled data for generalized category discovery. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3293–3303, 2025.

Lorenzo Caselli, Marco Mistretta, Simone Magistri, and Andrew Bagdanov. SpectralGCD: Spectral concept selection and cross-modal representation learning for generalized category discovery. In *Proceedings of International Conference on Learning Representations (ICLR)*, pp. 18162–18182, 2026.

Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23094–23104, 2024.

Qiyuan Dai, Hanzhuo Huang, Yu Wu, and Sibei Yang. Adaptive part learning for fine-grained generalized category discovery: A plug-and-play enhancement. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25444–25453, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *In Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Jizhou Han, Chenhao Ding, Yuhang He, Qiang Wang, Shaokun Wang, SongLin Dong, and Yihong Gong. Learning like humans: Analogical concept learning for generalized category discovery. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28734–28743, 2026.

Wei He, Xianghan Meng, Zhiyuan Huang, Xianbiao Qi, Rong Xiao, and Chun-Guang Li. Multimodal representation learning via semi-supervised rate reduction for generalized category discovery. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39637–39646, 2026.

Zhenqi He, Yuanpei Liu, and Kai Han. SEAL: Semantic-aware hierarchical learning for generalized category discovery. In *Proceedings of Conference on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 160765–160794, 2025.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python. 2020. doi: 10.5281/zenodo.1212303.

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 *Proceedings of International Conference on Machine Learning (ICML)*, pp. 4904–4916, 2021.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 554–561, 2013.

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

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Proceedings of Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of International Conference on Machine Learning (ICML)*, volume 162, pp. 12888–12900, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

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 *Proceedings of International Conference on Machine Learning (ICML)*, 2021.

Sarah Rastegar, Mohammadreza Salehi, Yuki M. Asano, Hazel Doughty, and Cees G. M. Snoek. Selex: Self-expertise in fine-grained generalized category discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset. *arXiv preprint arXiv:1906.05372*, 2019.

Luyao Tang, Kunze Huang, Chaoqi Chen, Yuxuan Yuan, Chenxin Li, Xiaotong Tu, Xinghao Ding, and Yue Huang. Dissecting generalized category discovery: Multiplex consensus under self-deconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 297–307, October 2025.

Luyao Tang, Kunze Huang, Chaoqi Chen, and Cheng Chen. Bures-isotropy alignment: Manifold learning of generalized category discovery. In *Proceedings of International Conference on Learning Representations (ICLR)*, pp. 55933–55957, 2026.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7492–7501, 2022.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Enguang Wang, Zhimao Peng, Zhengyuan Xie, Fei Yang, Xialei Liu, and Ming-Ming Cheng. Get: Unlocking the multi-modal potential of clip for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20296–20306, 2025.

Hongjun Wang, Sagar Vaze, and Kai Han. SPTNet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024.

Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16590–16600, 2023.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. FLAIR: VLM with fine-grained language-informed image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24884–24894, June 2025.

Qiyu Xu, Zhanxuan Hu, Yu Duan, Ercheng Pei, and Yonghang Tai. A hidden stumbling block in generalized category discovery: Distracted attention. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 405–414, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL `https://arxiv.org/abs/2505.09388`.

Muli Yang, Jie Yin, Yanan Gu, Cheng Deng, Hanwang Zhang, and Hongyuan Zhu. Consistent prompt tuning for generalized category discovery. *International Journal of Computer Vision (IJCV)*, 133(7):4014–4041, Jul 2025b.

Wei Zhang, Baopeng Zhang, Zhu Teng, Wenxin Luo, Junnan Zou, and Jianping Fan. Less attention is more: Prompt transformer for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30322–30331, June 2025.

Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Haiyang Zheng, Nan Pu, Yaqi Cai, Teng Long, Wenjing Li, Nicu Sebe, and Zhun Zhong. The devil is in gradient entanglement: Energy-aware gradient coordinator for robust generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3563–3573, June 2026.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

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