BEYOND PSEUDO-LABELS: VISION–LANGUAGE GUIDANCE FOR GENERALIZED CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized category discovery (GCD) aims to recognise known categories while discovering and clustering novel ones in a pool of unlabelled data. Previous methods rely on the visiual similarities between samples to form clusters, however this fail in the application where there is find discriminations between categories. On the other hand, vision language models (VLMs) carry rich semantic knowledge that could distinguish how fine grained categories can be distinguished from each other. hence complementing visual category discovery, yet existing approaches use this knowledge only through predicted labels or textual descriptions and a VLM's words are not always reliable exactly on the fine-grained novel classes. In this paper, We show that the internal semantic structure of a VLM provides useful supervision beyond its category predictions. We introduce BOLT, a framework that prompts a frozen VLM once per image for a coarse-to-fine taxonomy, reads its hidden state at every level, and uses this multi-level representation as a second view of the pseudo-label hierarchy of the vision encoder: the half of the contrastive embedding that GCD methods leave to an unsupervised loss is trained to the learner's hierarchical pseudo-labels from the image, from the VLM and from their average, and the VLM votes in the clustering that produces those pseudo-labels. BOLT never uses a VLM output as a label and leaves the baseline's objective and evaluation unchanged, so the discovery model keeps its visual discrimination while absorbing the VLM's knowledge. On three fine-grained benchmarks, added to five GCD methods (AL-GCD, SelEx, Hyp-SelEx, SimGCD and CMS) and four backbones (CLIP and DINOv1/v2/v3), BOLT improves AL-GCD by +5.7 points on average and by up to +9.3, and improves the DINOv1-based baselines by +11.5 on average and by up to 31 points; showing the benefits of the complementary semantic knowledg of VLMS.

1 INTRODUCTION

Supervised image classification usually assume that test images belong to a fixed set of classes seen during training; However in the open world a model must also recognise that an image belongs to a class it was never shown (Boult et al., 2019; Salehi et al., 2022; Troisemaine et al., 2023; He et al., 2025b). Generalized category discovery (GCD) (Vaze et al., 2022b; Cao et al., 2022) makes this concrete (Figure 2): given labelled images of a subset of classes and unlabelled images from those classes alongside images from novel ones, partition the unlabelled set into both known and novel categories. The most common approach adapts a self-supervised vision transformer (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2026) with contrastive and parametric objectives on the labelled and unlabelled data (Vaze et al., 2022b; Wen et al., 2023;

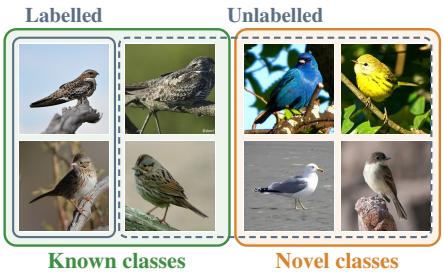
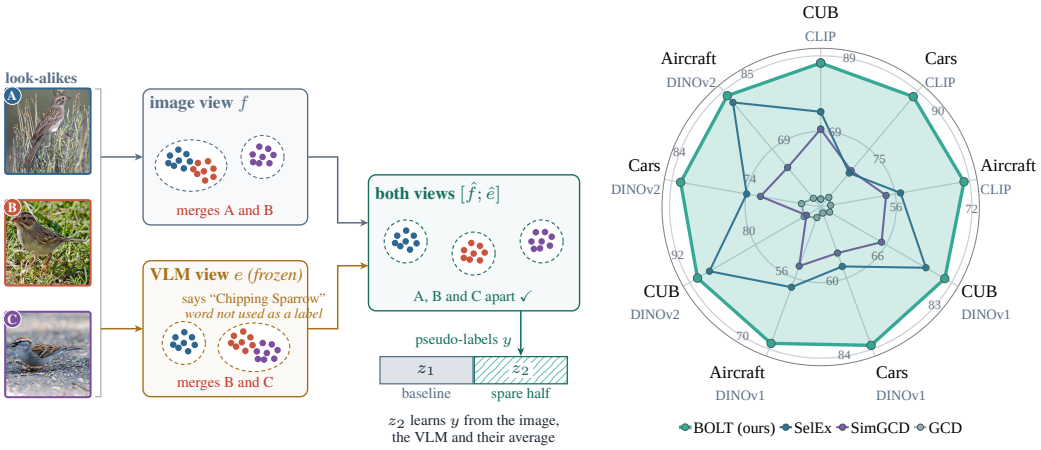


Figure 2: **Generalized category discovery.** Labelled images come from known classes; unlabelled images come from known and novel classes, and all of them must be clustered.

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) Overview of our approach

(b) Improvement over baselines for different backbones

Figure 1: **A VLM's words fail on novel classes; its hidden states are a second view.** (a) The image view and the VLM view merge different look-alikes; clustering both separates them, and the pseudo-labels train the spare half z_2 of the embedding. The VLM's words are never labels. (b) All accuracy of BOLT and of methods reported with all three backbones in Tables 1–2.

Zhang et al., 2023a; Wang et al., 2024; Rastegar et al., 2024; Choi et al., 2024), and recent work refines prototypes, parts, geometry and training signals (Liu et al., 2025d; Ma et al., 2025; Cendra & Han, 2026; He et al., 2025c; Tang et al., 2026; Liu et al., 2026c; Rastegar et al., 2026; Ye et al., 2026; Zheng et al., 2026a), also for discovery on the fly, under domain shift and in long-tailed data (Du et al., 2023; Wang et al., 2025b; Zhao & Han, 2025). With no labels for the novel classes, the learner supervises itself, so it clusters its own representation and trains on the resulting pseudo-labels and its errors concentrate where visual similarity misleads such as look-alike bird species, where the language component might highlight these differences, while visual similarity is often blind to them.

Recent GCD methods have used the language semantic benefits through CLIP (Radford et al., 2021): as retrieved text (Ouldnoughi et al., 2023), LLM-written descriptions of a tag lexicon (Zheng et al., 2024), pseudo-text embeddings synthesised from the image (Wang et al., 2025a). a filtered concept dictionary (Caselli et al., 2026), retrieved text aligned by rate reduction (He et al., 2026), tuned textual prompts (Solanki et al., 2024; Yang et al., 2025), a candidate label set (Liu et al., 2026a), LLM knowledge of the labelled class names (Banerjee & Biswas, 2025), or textual concepts generated by analogy to the labelled classes (Han et al., 2026). In all of them language enters as *short text or its embedding*, used as a label, a target or a feature, while the difference between fine-grained classes is not always represented well in the difference between word embeddings. Furthermore, since these language-based methods rely on the VLM's words, they are limited by the VLM's naming accuracy on novel classes and they might have a very high confidence in a wrong label since language componenet suggested it.

A generative vision language model (VLM) (Bai et al., 2025; Qwen Team, 2026; Zhu et al., 2025; Wang et al., 2025d; Comanici et al., 2025; Hong et al., 2026) can provide a much stronger representation, since we can ask for it to provid a taxonomy or different attirbutes and which coarse category an image belongs to so that we reduce the feature space for the possible categories. Its words, however, are unreliable and since these models do mistakes with high confidence that affect the downstream quality of the representations. VLMs underperform CLIP at classification although the class information is present in their latent space (Zhang et al., 2024), their fine-grained failures come from misaligned naming rather than missing evidence (He et al., 2025a), vocabulary-free naming of fine-grained classes remains far from supervised accuracy (Conti et al., 2023; Liu et al., 2024; Demidov et al., 2026), and longer textual reasoning lowers fine-grained accuracy (Zhu et al., 2026). On the novel classes that GCD must discover, a VLM is not a better labeller than the learner, so it cannot teach it labels.

What a VLM can offer instead is a second, largely independent view of the images, since its hidden states are usable embeddings (Jiang et al., 2025) and cluster images without labels (Wang et al., 2026) and GCD vision encoders have room for one. The contrastive objectives they inherit (Vaze

et al., 2022b; Rastegar et al., 2024; Wen et al., 2023; Zhang et al., 2023a; Wang et al., 2024) supervise only the leading coordinates of the embedding with labels and pseudo-labels; the remaining half is shaped by the instance-contrastive loss alone. BOLT (Figure 1) uses that spare half. It prompts a frozen VLM once per image for a fixed-key taxonomy from the broadest category down to one level finer than the dataset's labels, reads the hidden state at each level before the word is decoded, so the model never has to commit to a name. Then it trains the learner's own pseudo-label hierarchy from three views: the image, a learned projection of the VLM states, and their average. The VLM view also votes in the clustering that produces the pseudo-labels (Figure 3). Nothing the VLM says is a label; what it contributes is a similarity structure that is right, where the image features are wrong, often enough to move the clustering: since the clustering averages the two views' objectives, the VLM helps whenever it disagrees strongly enough with the wrong partitions the image view finds attractive.

Contributions.

1. We show that the hidden states of a generative VLM which reads once per taxonomy level before any word is decoded, can be a useful second view of a GCD vision encoder own pseudo-label hierarchy, while the VLM's words are unreliable on novel fine-grained classes.
2. We introduce BOLT, a plug-in objective that trains the half of a GCD embedding left to an unsupervised loss from three views, image, VLM and their average, and lets the frozen VLM vote in the pseudo-label clustering (Section 3). It adds $\approx$10M trainable parameters and one VLM pass per image, and leaves the baseline's objective and evaluation unchanged.
3. We add BOLT to five GCD methods on four backbones and three fine-grained benchmarks, with gains wherever the baseline is not already far above the VLM view, and show across VLM sizes that the gain is predicted by how well the VLM's states cluster the data rather than by its size or naming accuracy; the image model alone keeps the gain when the VLM can be removed at test time.

2 Related work

Generalized category discovery. Novel category discovery (NCD) transfers knowledge from labelled classes to cluster unlabelled images that belong to novel classes only (Hsu et al., 2018; Han et al., 2019; 2020; Zhao & Han, 2021; Zhong et al., 2021; Fini et al., 2021; Troisemaine et al., 2023). GCD (Vaze et al., 2022b) and open-world semi-supervised learning (Cao et al., 2022) remove this assumption: the unlabelled set mixes known and novel classes, the number of classes is given, and every unlabelled image must be assigned to a known class or to a newly discovered one, with accuracy measured after one Hungarian matching over all classes. The original GCD method fine-tunes a DINO ViT with supervised and self-supervised contrastive losses and clusters its features with semi-supervised k-means. Later non-parametric methods refine the representation or the clustering (Fei et al., 2022; Zhang et al., 2023a; Pu et al., 2023; Choi et al., 2024; Otholt et al., 2024; Zhao et al., 2023), parametric methods train a classifier over all classes by self-distillation and correct its biases (Wen et al., 2023; Vaze et al., 2023; Cao et al., 2024; Lin et al., 2024; Liu & Han, 2025; Guo et al., 2025), and much recent work changes how the backbone is adapted, with prompts and adapters (Wang et al., 2024; Ye et al., 2026), token and attention refinement (Xu et al., 2025; Yang et al., 2026), part-level cues (Dai et al., 2025; Cendra & Han, 2026), geometry (Liu et al., 2025d; Tang et al., 2026), distillation from a closed-set teacher (Liu et al., 2026c) or frequency cues (Rastegar et al., 2026). Related settings discover categories on the fly (Du et al., 2023), incrementally (Zhao & Mac Aodha, 2023) or under domain shift (Wang et al., 2025b). Closest to us are methods that model the hierarchy of categories (Rastegar et al., 2023; 2024; He et al., 2025c; Zheng et al., 2026b): SelEx builds progressively coarser pseudo-label levels and supervises nested prefixes of the embedding with them. All of these derive category structure from the images alone; BOLT plugs into such an objective and adds a second view of that structure from a VLM.

Language in category discovery. CLIP-GCD (Ouldnoughi et al., 2023) retrieves corpus text for joint image-text clustering, TextGCD (Zheng et al., 2024) co-teaches a text and an image branch on LLM-written tag descriptions, GET (Wang et al., 2025a) synthesises pseudo-text embeddings, GraphVL (Solanki et al., 2024) and CPT (Yang et al., 2025) tune textual prompts, Spectral-GCD

(Caselli et al., 2026) selects concepts from a dictionary, AL-GCD (Han et al., 2026) generates analogical textual concepts, SSR^2-GCD (He et al., 2026) aligns modalities by rate reduction, SECOS (Liu et al., 2026a) predicts labels from a candidate text set, and SDBA (Qu et al., 2026) adapts both CLIP branches. Beyond GCD, (M)LLMs write class descriptions (Menon & Vondrick, 2023; Pratt et al., 2023), name fine-grained classes without a vocabulary (Conti et al., 2023; Liu et al., 2024; Demidov et al., 2025; 2026), cluster images through generated text or concepts (Kwon et al., 2024; Li et al., 2024; Liu et al., 2026b; Stephan et al., 2024; Zhong et al., 2026), and guide text clustering and text GCD (Zhang et al., 2023b; An et al., 2024). In all of these the (M)LLM contributes through the text it produces, injected as a feature or a target. BOLT uses a generative VLM's hidden states, never its words as labels, and only as a second view of the learner's own pseudo-labels.

Hidden states of multimodal LLMs, and multi-view learning. Generative VLMs connect a vision encoder to a language model (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Bai et al., 2025; Qwen Team, 2026). Prompted last-token states of (M)LLMs are usable embeddings (BehnamGhader et al., 2024; Jiang et al., 2024a;b; 2025), representations of different models and modalities converge (Huh et al., 2024), intermediate layers can beat the last (Skean et al., 2025), and embeddings read after the model has reasoned improve retrieval (Cui et al., 2026; Lan et al., 2026; Liu et al., 2025b); ESMC (Wang et al., 2026) clusters images with prompt-conditioned states of a frozen MLLM, without labelled data, and LLM attention can be clustered in context (Wang et al., 2025e). Probing shows that the language model often fails to use category information that its states contain (Fu et al., 2025; Liu et al., 2025a). Such states are strongly anisotropic, which all-but-the-top post-processing corrects (Mu & Viswanath, 2018). Training two views of the same data to agree on labels is co-training (Blum & Mitchell, 1998; Dasgupta et al., 2001); supervising nested prefixes of an embedding is the Matryoshka idea (Kusupati et al., 2022) that SelEx applies to pseudo-label levels. To our knowledge, no prior GCD method reads the hidden states of a generative VLM, and none uses an external model as a view of the learner's own partition rather than as a source of labels or features.

3 METHOD

3.1 PRELIMINARIES: GENERALIZED CATEGORY DISCOVERY

Problem Definition. GCD assumes that model receives a labelled set $\mathcal{D}_l = \{(x_i, y_i)\}$ with $y_i \in \mathcal{Y}_l$ and an unlabelled set $\mathcal{D}_u = \{x_j\}$ whose images belong to the known classes $\mathcal{Y}_l$ *or* to novel classes $\mathcal{Y}_n$, with $\mathcal{Y}_l \cap \mathcal{Y}_n = \emptyset$. The task is to assign every unlabelled image to one of the $|\mathcal{Y}_l| + |\mathcal{Y}_n|$ classes; the number of classes is known, as in the methods we build on. Accuracy is measured on $\mathcal{D}_u$ after one Hungarian matching of predicted clusters to classes over all classes, and reported for all, known and novel classes.

Contrastive GCD with a pseudo-label hierarchy. The methods we build on map an image to a feature f – the DINO (Caron et al., 2021) or CLIP (Radford et al., 2021) image feature, for AL-GCD (Han et al., 2026) mixed with an analogical text feature and a unit-norm projection $z = g(f) \in \mathbb{R}^D$. These methods train the vision encoder with an instance-contrastive loss $\mathcal{L}_u$ on two augmented views of every image and a supervised contrastive loss $\mathcal{L}_{lab}$ (Khosla et al., 2020) on the labelled images (Vaze et al., 2022b), and cluster with semi-supervised k-means, the labelled images pinned to their classes. SelEx (Rastegar et al., 2024), which AL-GCD and Hyp-SelEx (Liu et al., 2025d) inherit, runs this clustering every epoch on the training images and builds pseudo-label levels $y^{(0)}, \ldots, y^{(L-1)}$, halving the numbers of known and novel clusters from one level to the next, and supervises nested prefixes of z with them:

$$\mathcal{L}_{\text{base}} = (1-\lambda)\,\mathcal{L}_u + \frac{\lambda}{2}\Big[\mathcal{L}_{lab} + \sum_{i=0}^{L-1} 2^{-(i+1)}\,\mathcal{S}\big(z^{:m_i}, y^{(i)}\big)\Big], \qquad m_i = D/2^{i+1}, \tag{1}$$

where $\mathcal{S}(u, y)$ is the supervised contrastive loss with labels y and $u^{:m}$ the first m coordinates of u. Every pseudo-label level acts on the first half $z_1 = z^{:D/2}$; the second half z_2 is shaped by $\mathcal{L}_u$ alone. Our method BOLT can be added to any of these methods and uses the spare half z_2 (Figure 3). The process has two stages. *Before training*, a frozen VLM describes every image once as a coarse-to-fine taxonomy, and its hidden state at each level is read and compressed into a vector

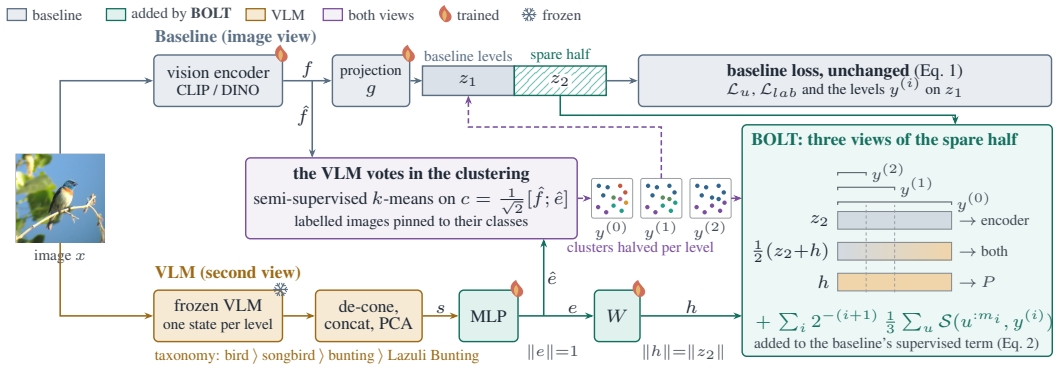


Figure 3: **BOLT.** The baseline maps an image to a feature f and an embedding $z = [z_1; z_2]$; its loss is unchanged and its pseudo-label levels act on z_1. A frozen VLM answers a taxonomy prompt once per image, and its per-level hidden states, with their mean and top directions removed, are projected to h. The VLM votes in the clustering that produces the levels (k-means on $[\hat{f}; \hat{e}]$), and the spare half z_2 is trained to the same levels from three views, z_2, h and their average on nested prefixes of increasing length. Flame: trained; snowflake: frozen.

s per image (Section 3.2); this is done once and needs no labels. *During training*, at every epoch, (i) the baseline's pseudo-label hierarchy is computed by clustering the image feature together with the VLM embedding, so the VLM votes in the clustering (Section 3.4); (ii) for every minibatch, the baseline's own loss acts on z_1 exactly as before, while z_2 is trained to the same pseudo-label levels from three views: the image view z_2, the VLM view h obtained by a small trainable projection of s, and their average (Section 3.3). At test time the baseline's own evaluation is used; the VLM can be kept or dropped.

3.2 Reading the VLM

Each image is shown once to a frozen VLM with a prompt that asks for a fixed-key taxonomy, coarse to fine (for CUB: `domain, class, group, family, genus, species, subtype`), writing `unknown` for any level the image does not support. The prompt names no novel class; on CUB it contains one known-class name, which GCD permits (Appendix J). The answer is then fed back with the image and prompt in one teacher-forced pass, and for each level we read the last-layer state at the last token of that level's value. The states of different images point in nearly the same direction (mean cosine 0.82), so for each level we remove the mean and the top eight principal directions (Mu & Viswanath, 2018), normalise, concatenate across levels and reduce by PCA to $s \in \mathbb{R}^{512}$. All of this is computed once and uses no labels.

3.3 Three views of the spare half

A trainable projection $P(s) = W\,\mathrm{normalize}(\mathrm{MLP}(s))$ maps the VLM representation to $h \in \mathbb{R}^{D/2}$, rescaled to the norm of z_2 (no gradient through the scale). BOLT adds to the baseline's supervised term

$$\sum_{i=0}^{L-1} 2^{-(i+1)}\,\frac{1}{3}\sum_{u\in\{z_2,\,h,\,\frac{1}{2}(z_2+h)\}} \mathcal{S}\big(u^{:m_i},\,y^{(i)}\big), \tag{2}$$

so the second half receives the same hierarchy as the first, split equally between the image view, the VLM view and their average. This way we ensure that while we learn a projection from the VLM view to image encoder pseudolabels, we also inherit the semantic knowledge that helps the vision encoder to align with semantic knowledge of VLM thorugh projecting their average to the psudodolables as wel. The terms train different parameters: $\mathcal{S}(z_2)$ the image model, $\mathcal{S}(h)$ only P, and the average both. The average's squared distance expands exactly to $\frac{1}{4}(\|z_{2a} - z_{2b}\|^2 + \|h_a -$

$h_b\|^2$) plus a cross-modal term that, for a pseudo-positive pair, pulls up $\langle z_{2a}, h_b\rangle + \langle h_a, z_{2b}\rangle$: a co-training signal whose positives are the learner's own pseudo-label classes.

3.4 THE VLM VOTES IN THE CLUSTERING

The pseudo-labels are computed on $c = \frac{1}{\sqrt{2}}[\hat{f}; \hat{e}]$ instead of $\hat{f}$, where $e = \mathrm{normalize}(\mathrm{MLP}(s))$ is the bottleneck of P. Since $\cos(c_a, c_b) = \frac{1}{2}\big(\cos(f_a, f_b) + \cos(e_a, e_b)\big)$, the k-means objective on c is exactly the average of the objectives on the two views, so the VLM overturns a wrong partition whenever it disfavours it more than the image view favours it. AL-GCD and SelEx evaluate on the same c; when either modality is unavailable at test time, f or e alone can be clustered (Section A). For SimGCD and CMS, whose evaluation uses their own head or embedding, the VLM acts through training only. Algorithm 1 in Appendix B summarises the procedure.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on three fine-grained benchmarks of the Semantic Shift Benchmark (Vaze et al., 2022a): CUB-200 (Wah et al., 2011), Stanford Cars (Krause et al., 2013) and FGVC-Aircraft (Maji et al., 2013), with the standard SSB class split in which half of the classes are known and the other ones are unknown. Following Vaze et al. (2022b), half of the images of the known classes are labelled and the remaining images of all classes form the unlabelled set.

Evaluation. Similar to other GCD methods, we report the accuracy on the unlabelled training images for all, known and novel classes. Each baseline is evaluated with its own protocol at the end of training, agglomerative clustering for AL-GCD, semi-supervised k-means for SelEx and Hyp-SelEx, the classifier head for SimGCD, and mean-shift followed by agglomerative clustering for CMS.

Baselines. We add BOLT to AL-GCD (Han et al., 2026) which uses (CLIP ViT-B/16) as the backbone, SelEx (Rastegar et al., 2024) (DINOv1, DINOv2 and DINOv3 (Siméoni et al., 2026) ViT-B), Hyp-SelEx (Liu et al., 2025d) (DINOv1 and DINOv2), SimGCD (Wen et al., 2023) and CMS (Choi et al., 2024) (DINOv1 ViT-B/16). Every baseline is reproduced with its official code and released hyperparameters and BOLT keeps the baseline's supervised weight ($\lambda = 0.35$ for AL-GCD).

Implementation details. The VLM is Qwen3.5-9B unless stated otherwise, prompted once per training image with the taxonomy prompts of Appendix J. For each level we remove the mean and the top eight principal directions of its states, normalise, concatenate and reduce by PCA to 512 dimensions. The projection P is an MLP ($512 \to 2048 \to 256$, bottleneck e) followed by a linear map W to the $D/2$ coordinates of z_2, $\approx$10M parameters in total, trained with the baseline's optimiser. All other hyperparameters (epochs, learning rate, fine-tuned blocks, number of levels) are the same as the baseline's.

4.2 COMPARISON WITH THE STATE OF THE ART

Tables 1 and 2 add BOLT to five baselines on four different backbones. Each baseline is trained using its released code (rows †), and the shaded row below it changes only BOLT; the last row of each block is the gain averaged over that block's baselines.

CLIP. On AL-GCD, BOLT gains +5.7 All on average over the three fine-grained benchmarks (known +3.8, novel +6.5). The gains are largest where the image model is weakest relative to the VLM: Stanford Cars (+9.3; novel classes +12.8), where raw CLIP image features cluster car models at 6%, and FGVC-Aircraft (+6.0; novel classes +6.8), where they cluster aircraft variants at 20%. Among CLIP methods, AL-GCD with BOLT has the highest accuracy on CUB and FGVC-Aircraft and the highest fine-grained average (82.2, against 77.1 for Spectral-GCD); on Stanford Cars it is just below Spectral-GCD and SSR^2-GCD, which retrieves text with a larger CLIP-H/14 encoder (88.6 vs 89.1 and 89.2). The trend follows how well CLIP separates each benchmark on its own: the

Table 1: **Fine-grained benchmarks, CLIP and DINOv1 backbones**, All / Known / Novel (%). Shaded: the baseline above it with BOLT and nothing else changed; the last row of each block averages the gain of BOLT over that block's baselines. Bold and underline: best and second best per backbone.

	Method	Venue	CUB			Stanford Cars			FGVC-Aircraft			Average		
			All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
CLIP	GCD§ (Vaze et al., 2022b)	*CVPR22*	51.1	56.2	48.6	62.5	73.9	57.0	41.2	43.0	40.2	51.6	57.7	48.6
CLIP	SimGCD§ (Wen et al., 2023)	*ICCV23*	69.6	75.8	66.5	69.4	82.2	62.9	53.5	57.9	51.7	64.2	72.0	60.4
CLIP	CMS§ (Choi et al., 2024)	*CVPR24*	65.8	75.3	61.1	77.9	89.0	72.6	50.3	59.1	45.9	64.7	74.5	59.9
CLIP	SelEx§ (Rastegar et al., 2024)	*ECCV24*	74.2	69.5	76.5	68.8	68.4	69.0	56.7	58.2	56.0	66.6	65.4	67.2
CLIP	TextGCD (Zheng et al., 2024)	*ECCV24*	76.6	80.6	74.7	86.9	87.4	86.7	50.8	44.9	53.8	71.4	71.0	71.7
CLIP	GET (Wang et al., 2025a)	*CVPR25*	77.0	78.1	76.4	78.5	86.8	74.5	58.9	59.6	58.5	71.5	74.8	69.8
CLIP	Spectral-GCD (Caselli et al., 2026)	*ICLR26*	79.2	80.4	78.5	89.1	92.6	**87.4**	63.0	66.1	61.4	77.1	79.7	75.8
CLIP	SSR^2-GCD (He et al., 2026)	*CVPR26*	78.3	78.5	78.2	**89.2**	93.1	87.3	–	–	–	–	–	–
CLIP	AL-GCD† (Han et al., 2026)	*CVPR26*	85.4	76.7	**89.7**	79.3	92.2	73.1	64.8	62.0	66.2	76.5	77.0	76.3
CLIP	+ **BOLT**		**87.1**	**82.0**	**89.7**	88.6	**94.2**	85.9	**70.8**	**66.3**	**73.0**	**82.2**	**80.8**	**82.9**
CLIP	*Average gain of + BOLT*		+1.7	+5.3	0.0	+9.3	+1.9	+12.8	+6.0	+4.3	+6.8	+5.7	+3.8	+6.5
DINOv1	GCD (Vaze et al., 2022b)	*CVPR22*	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	45.1	51.8	41.8
DINOv1	PromptCAL (Zhang et al., 2023a)	*CVPR23*	62.9	64.4	62.1	50.2	70.1	40.6	52.2	52.2	52.3	55.1	62.2	51.7
DINOv1	μGCD (Vaze et al., 2023)	*NeurIPS23*	65.7	68.0	64.6	56.5	68.1	50.9	53.8	55.4	53.0	58.7	63.8	56.2
DINOv1	InfoSieve (Rastegar et al., 2023)	*NeurIPS23*	69.4	77.9	65.2	55.7	74.8	46.4	56.3	63.7	52.5	60.5	72.1	54.7
DINOv1	SPTNet (Wang et al., 2024)	*ICLR24*	65.8	68.8	65.1	59.0	79.2	49.3	59.3	61.8	58.1	61.4	69.9	57.5
DINOv1	LegoGCD (Cao et al., 2024)	*CVPR24*	63.8	71.9	59.8	57.3	75.7	48.4	55.0	61.5	51.7	58.7	69.7	53.3
DINOv1	PAL-GCD (Wang et al., 2025c)	*AAAI25*	72.6	75.2	71.2	72.2	83.4	66.8	58.8	64.5	56.0	67.9	74.4	64.7
DINOv1	DebGCD (Liu & Han, 2025)	*ICLR25*	66.3	71.8	63.5	65.3	81.6	57.4	61.7	63.9	60.6	64.4	72.4	60.5
DINOv1	CloSeR (Liu et al., 2026c)	*ECCV26*	**82.8**	80.3	**84.1**	66.5	83.4	58.3	66.6	68.1	65.8	72.0	77.3	69.4
DINOv1	FourEx (Rastegar et al., 2026)	*ECCV26*	80.2	75.1	82.7	59.2	76.9	50.6	65.9	67.9	64.9	68.4	73.3	66.1
DINOv1	SimGCD† (Wen et al., 2023)	*ICCV23*	64.8	67.6	63.4	52.2	69.7	43.8	54.2	56.0	53.3	57.1	64.4	53.5
DINOv1	+ **BOLT**		71.0	76.2	68.3	66.5	87.1	56.5	58.2	66.5	54.1	65.2	76.6	59.7
DINOv1	CMS† (Choi et al., 2024)	*CVPR24*	67.4	73.4	64.5	50.3	69.3	41.1	51.0	55.6	48.7	56.2	66.1	51.4
DINOv1	+ **BOLT**		77.5	**82.2**	75.1	81.6	**92.3**	76.5	56.2	63.2	52.7	71.8	79.3	68.1
DINOv1	SelEx† (Rastegar et al., 2024)	*ECCV24*	76.3	76.1	76.4	56.7	77.3	46.7	58.2	65.1	54.8	63.8	72.8	59.3
DINOv1	+ **BOLT**		81.2	81.0	81.3	**82.8**	91.6	**78.5**	**68.9**	**72.5**	**67.2**	**77.6**	**81.7**	**75.7**
DINOv1	Hyp-SelEx† (Liu et al., 2025d)	*CVPR25*	78.6	73.1	81.3	57.4	73.0	49.9	58.8	60.6	57.9	64.9	68.9	63.0
DINOv1	+ **BOLT**		81.5	77.0	83.8	72.0	87.6	64.5	66.2	68.4	65.1	73.3	77.7	71.1
DINOv1	*Average gain of + BOLT*		+6.0	+6.6	+5.7	+21.6	+17.3	+23.6	+6.8	+8.3	+6.1	+11.5	+10.7	+11.8

† Reproduced using the official released code.
§ Reported from AL-GCD (Han et al., 2026).

gain is smallest on CUB (+1.7), where CLIP features already separate bird species, and largest on Cars (+9.3).

DINOv1. BOLT is the same on every baseline (Section 3); SelEx and Hyp-SelEx receive it unchanged, and for SimGCD and CMS, which have no pseudo-label hierarchy, the levels are built each epoch by semi-supervised k-means on $[\hat{f}; \hat{e}]$ exactly as SelEx builds them (Appendix C). All four baselines gain on every benchmark, by +11.5 All on average (known +10.7, novel +11.8), most on Stanford Cars, where the DINOv1 baselines cluster car models at 50–57% and BOLT lifts them to 66–83%: SelEx from 56.7 to 82.8, CMS from 50.3 to 81.6. SelEx with BOLT averages 77.6 over the three benchmarks, above every published DINOv1 method. On CUB the image model alone, clustered without the VLM, keeps the whole gain of SelEx (+5.7 against the baseline's own image feature), so the knowledge has moved into the DINO backbone. Across the four baselines the gain is larger the weaker the baseline, and largest on Stanford Cars, the benchmark on which DINOv1 features are weakest.

DINOv2 and DINOv3. With the stronger backbones BOLT still gains on Stanford Cars (+9.2 and +8.7 for SelEx) and FGVC-Aircraft (+1.5, +2.8 and +2.9), but not on CUB (-1.3, +0.1 and +0.4), where these baselines cluster far above the VLM view, nor for Hyp-SelEx on Stanford Cars (-0.8), which is already at 83.6. Overall the gain shrinks as the backbone gets stronger (+11.5 with DINOv1, +2.0 with DINOv2 and +3.8 with DINOv3) and concentrates on the benchmarks where the backbone is weakest (Appendix A). On DINOv3, SelEx with BOLT is above PartCo-SelEx (Cendra & Han, 2026), which adds part-level correspondences to SelEx, on Stanford Cars (84.7 vs 81.7) and below it on CUB (84.2 vs 86.1) and FGVC-Aircraft (75.2 vs 76.9); the two add different information, visual part correspondences and a language view to the same baseline.

Table 2: **Fine-grained benchmarks, DINOv2 and DINOv3 backbones**, All / Known / Novel (%). Notation as in Table 1.

	Method	Venue	CUB			Stanford Cars			FGVC-Aircraft			Average		
			All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
DINOv2	GCD* (Vaze et al., 2022b)	*CVPR22*	71.9	71.2	72.3	65.7	67.8	64.7	55.4	47.9	59.2	64.3	62.3	65.4
	CiPR (Hao et al., 2024)	*TMLR24*	78.3	73.4	80.8	66.7	77.0	61.8	–	–	–	–	–	–
	SimGCD* (Wen et al., 2023)	*ICCV23*	71.5	78.1	68.3	71.5	81.9	66.6	63.9	69.9	60.9	69.0	76.6	65.3
	μGCD* (Vaze et al., 2023)	*NeurIPS23*	74.0	75.9	73.1	76.1	91.0	68.9	66.3	68.7	65.1	72.1	78.5	69.0
	FlipClass (Lin et al., 2024)	*NeurIPS24*	79.3	80.7	78.5	78.0	88.0	73.2	71.1	75.1	69.1	76.1	81.3	73.6
	DebGCD (Liu & Han, 2025)	*ICLR25*	77.5	80.8	75.8	75.4	87.7	69.5	71.9	76.0	69.8	74.9	81.5	71.7
	RLCD (Liu et al., 2025c)	*ICML25*	78.7	79.5	78.3	79.5	91.8	73.5	72.6	77.3	70.3	76.9	82.9	74.0
	ConGCD (Tang et al., 2025)	*ICCV25*	86.3	87.4	85.8	79.8	93.1	73.3	81.7	83.3	81.0	82.6	87.9	80.1
	SEAL (He et al., 2025c)	*NeurIPS25*	76.7	78.3	75.9	77.7	88.7	72.4	74.6	73.2	75.3	76.3	80.1	74.5
	PartCo-SelEx (Cendra & Han, 2026)	*ICML26*	**90.6**	84.5	93.2	82.5	91.8	78.0	83.4	83.6	83.3	85.5	86.6	84.8
	CloSeR+SelEx (Liu et al., 2026c)	*ECCV26*	89.8	86.5	91.5	83.1	**94.6**	77.6	82.1	81.2	82.6	85.0	87.4	83.9
	FourEx (Rastegar et al., 2026)	*ECCV26*	87.8	86.3	88.6	80.1	94.3	73.2	81.5	84.7	80.0	83.1	88.4	80.6
	NN-GCD (Ji et al., 2026)	*TCSVT26*	83.1	82.5	83.4	80.4	88.9	76.3	75.6	78.2	74.3	79.7	83.2	78.0
	SelEx† (Rastegar et al., 2024)	*ECCV24*	88.5	85.2	90.2	73.4	88.8	66.0	81.9	83.0	81.4	81.3	85.6	79.2
	+ **BOLT**		87.2	**87.5**	87.1	82.7	93.7	77.4	83.4	**84.9**	82.7	84.4	**88.7**	82.4
	Hyp-SelEx† (Liu et al., 2025d)	*CVPR25*	90.2	81.5	94.6	**83.6**	90.3	**80.3**	80.8	81.2	80.7	84.9	84.3	85.2
	+ **BOLT**		**90.6**	81.8	**95.0**	82.8	89.3	79.6	**83.8**	83.9	**83.7**	**85.7**	85.0	**86.1**
	Average gain of + BOLT		−0.5	+1.3	−1.3	+4.2	+2.0	+5.3	+2.2	+2.3	+2.2	+2.0	+1.9	+2.0
DINOv3	SimGCD° (Wen et al., 2023)	*ICCV23*	75.9	83.8	72.0	73.9	79.4	71.3	71.4	**84.4**	65.0	73.7	82.5	69.4
	PartCo-SelEx (Cendra & Han, 2026)	*ICML26*	**86.1**	84.4	**87.0**	81.7	92.1	76.6	**76.9**	82.2	**74.2**	**81.6**	**86.2**	**79.3**
	SelEx† (Rastegar et al., 2024)	*ECCV24*	84.1	83.0	84.7	76.0	88.8	69.9	72.4	75.4	70.9	77.5	82.4	75.2
	+ **BOLT**		84.2	**86.1**	83.2	**84.7**	**93.8**	**80.3**	75.2	77.6	74.0	81.4	85.9	79.2
	Average gain of + BOLT		+0.1	+3.1	−1.5	+8.7	+5.1	+10.4	+2.8	+2.2	+3.1	+3.8	+3.5	+4.0

† Reproduced using the official released code.
* Reported from Vaze et al. (2023).
° Reported from PartCo (Cendra & Han, 2026).

Table 3: **Components of BOLT**, added one at a time to AL-GCD (CLIP ViT-B/16), All / Known / Novel (%), $\lambda = 0.35$ in every row. *levels on* z_2: the baseline's pseudo-label levels on the spare half, image view only; *VLM view*: the VLM view h and the average; *VLM vote*: the VLM in the clustering; *nested*: nested levels on the tail instead of one level. The first and the shaded row are the AL-GCD rows of Table 1. Bold: best per column.

Components					CUB			Stanford Cars			FGVC-Aircraft		
Method	levels on z_2	VLM view	VLM vote	nested	All	Known	Novel	All	Known	Novel	All	Known	Novel
AL-GCD	×	×	×	×	85.4	76.7	**89.7**	79.3	92.2	73.1	64.8	62.0	66.2
	✓	×	×	×	81.8	83.3	81.0	79.4	92.0	73.3	66.8	64.7	67.9
	✓	✓	×	×	83.5	**83.5**	83.5	78.9	92.7	72.2	65.3	61.9	67.0
	✓	✓	✓	×	83.0	81.4	83.9	86.1	**94.2**	82.2	67.9	**71.2**	66.2
BOLT	✓	✓	✓	✓	**87.1**	82.0	**89.7**	**88.6**	**94.2**	**85.9**	**70.8**	66.3	**73.0**

4.3 Ablation Studies

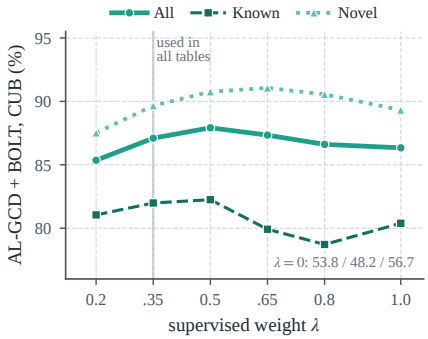
All ablations add BOLT to AL-GCD (CLIP ViT-B/16) in the configuration of Table 1.

Components. Supervising the spare half with the baseline's pseudo-labels but no VLM does nothing on Cars, hurts CUB and stays within run-to-run variation on FGVC-Aircraft (Table 3): the gain is the VLM's, not that of extra supervision. On Cars the VLM helps once it also votes in the clustering; on CUB and FGVC-Aircraft the nested levels are needed as well – on Aircraft the vote alone lifts the known classes by 9 points and the novel classes not at all, and the nested levels bring the novel classes up by 7.

How the VLM is read. One state per level matters: reading a single state at the end of the answer instead loses 2.0 points on CUB and 6.5 on Cars, almost all of it on the novel classes (Figure 4a). The useful levels are the coarse and intermediate ones. The coarse levels alone match all levels (88.0 on CUB, 89.1 on Cars), while the fine levels alone lose 5.2 and 4.4 points and fall below AL-GCD on CUB: at the species or trim level the VLM commits to a name it often gets wrong (Appendix J), whereas the coarser states have integrated the image evidence into a category decision without that commitment. Reading the state just before each value stays within run-to-run variation of BOLT (Figure 4a); training the VLM view on the full embedding instead of the spare half does too on CUB

Variant	CUB All	CUB Known	CUB Novel	Stanford Cars All	Stanford Cars Known	Stanford Cars Novel
BOLT	87.1	**82.0**	89.7	88.6	94.2	85.9
h on the full embedding	87.3	79.0	91.5	88.5	**94.6**	85.5
state before each value	**88.1**	80.9	91.8	87.6	94.3	84.4
one state per image	85.1	81.8	86.8	82.1	**94.6**	76.0
coarse levels only	88.0	79.9	**92.0**	**89.1**	94.0	**86.7**
fine levels only	81.9	79.7	83.1	84.2	93.9	79.5

(a) Variants of BOLT



(b) Supervised weight λ

Figure 4: **Ablations of AL-GCD + BOLT**, All / Known / Novel (%). (a) One design choice replaced: the VLM view on the full embedding instead of z_2; the state just before each value instead of its last token; one state per image instead of one per level; only the coarse (CUB `domain–family`, Cars `domain–make`) or fine levels. Bold: best per column. (b) CUB against the supervised weight; every table uses $\lambda = 0.35$ (vertical line), and without any supervised term ($\lambda = 0$) All collapses to 53.8.

and Cars but loses 2.6 points on FGVC-Aircraft, where the baseline's own levels on z_1 are then no longer separated from the VLM's.

Supervised weight. Because BOLT adds supervised terms, a natural question is whether the gain is a matter of supervised weight. Every comparison in this paper uses the baseline's own $\lambda = 0.35$ for both arms, and BOLT is stable around it (Figure 4b): it stays within 1.6 points of its best for $\lambda \in [0.35, 1]$ and is best at 0.5 (BOLT at $\lambda = 0.5$ is in Appendix D). The baseline itself does not profit from a different weight: AL-GCD is best at its default and loses 1.7–3.6 points at every other λ from 0.2 to 1, and at every λ BOLT is above AL-GCD trained with the same λ, by 1.7–4.6 points (Table 5).

Further analyses (missing modality at test time, pseudo-label quality, VLM size, and when the VLM helps) are in Appendix A. Success and failure cases, attention maps and a t-SNE of the learned features are in Appendix K.

5 Conclusion

A VLM is too unreliable to label novel classes but reliable enough to be a second view of the learner's own pseudo-labels. BOLT supervises the spare half of a GCD embedding from the image, the VLM and their average, lets the VLM vote in the clustering, and improves five GCD methods wherever the baseline is neither at its ceiling nor far above the VLM view. What matters in a VLM is how well its states cluster the data, which is measurable before training and not implied by size or naming accuracy.

Limitations. BOLT does not help a baseline that already clusters a dataset far better than the VLM view (CUB with DINOv2 or DINOv3, where novel classes lose) or at its ceiling (Stanford Cars with Hyp-SelEx on DINOv2), and we evaluate on fine-grained benchmarks only. Every result is a single run, paired with its baseline on the same hardware, and differences below about two points are within run-to-run variation (Appendix F). The number of classes is assumed known, as in the baselines; the VLM's pretraining data may include these benchmarks; and one VLM pass per training image is required.

AI use statement

In this work, we used generative AI tools for experiment design, the implementation of the training and analysis code, the interpretation of results. We have not used generative AI tools to create or annotate data; the vision–language models that are part of BOLT are described in Section 3.

Additionally, we used generative AI tools for literature search, for checking the numbers copied into the comparison tables against the original papers, and for drafting and editing the text. We have reviewed all AI-assisted work: every number in the paper is generated from the training logs by a script and was checked against the logs; the code was verified by tests showing that each baseline is unchanged when BOLT is disabled; and related-work claims were checked against the cited papers. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

All datasets are public benchmarks used under their licences; the VLMs are open-weight models used frozen. BOLT inherits any biases of the VLM's web-scale pretraining; it never uses the VLM's words as labels, which limits the effect of naming biases, but the taxonomies it writes may still reflect them.

Reproducibility statement

Section 3 and the setup paragraph of Section 4 specify BOLT and its hyperparameters; Appendix C describes how it is added to each baseline, Appendix J gives the prompts and the leak check, and Appendix F the evaluation protocol. The code, the prompts, the cached VLM states and the commands of every run will be released; all numbers in the paper are generated from the run logs by a script that is part of the release. The alternative level reading of Appendix I is a one-line switch in the released code.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23716–23736, 2022.

Wenbin An, Wenkai Shi, Feng Tian, Haonan Lin, QianYing Wang, Yaqiang Wu, Mingxiang Cai, Luyan Wang, Yan Chen, Haiping Zhu, and Ping Chen. Generalized category discovery with large language models in the loop. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8653–8665. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.512.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Anwesha Banerjee and Soma Biswas. Language-assisted feature representation and lightweight active learning for on-the-fly category discovery. *Transactions on Machine Learning Research*, 2025.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. In *Conference on Language Modeling (COLM)*, 2024.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Conference on Computational Learning Theory (COLT)*, 1998.

Terrance E Boult, Steve Cruz, Akshay Raj Dhamija, Manuel Gunther, James Henrydoss, and Walter J Scheirer. Learning and the unknown: Surveying steps toward open world recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9801–9807, 2019.

Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *Proceedings of the International Conference on Learning Representations*, 2022.

Xinzi Cao, Xiawu Zheng, Guanhong Wang, Weijiang Yu, Yunhang Shen, Ke Li, Yutong Lu, and Yonghong Tian. Solving the catastrophic forgetting problem in generalized category discovery.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16880–16889, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Lorenzo Caselli, Marco Mistretta, Simone Magistri, and Andrew D. Bagdanov. SpectralGCD: Spectral concept selection and cross-modal representation learning for generalized category discovery. In *The Fourteenth International Conference on Learning Representations*, 2026.

Fernando Julio Cendra and Kai Han. PartCo: Part-level correspondence priors enhance category discovery. In *International Conference on Machine Learning (ICML)*, 2026.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. In *Advances in Neural Information Processing Systems*, 2023.

Xuanming Cui, Jianpeng Cheng, Hong-You Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Taipeng Tian, Qi Guo, Ser-Nam Lim, Aashu Singh, and Xiangjun Fan. Think then embed: Generative context improves multimodal embedding. In *International Conference on Learning Representations (ICLR)*, 2026.

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 the Computer Vision and Pattern Recognition Conference*, pp. 25444–25453, 2025.

Sanjoy Dasgupta, Michael L. Littman, and David McAllester. PAC generalization bounds for co-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2001.

Dmitry Demidov, Muhammad Zaigham Zaheer, Omkar Thawakar, Salman Khan, and Fahad Shahbaz Khan. Vocabulary-free fine-grained visual recognition via enriched contextually grounded vision-language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 4275–4284, 2025.

Dmitry Demidov, Muhammad Zaigham Zaheer, Zongyan Han, Omkar Thawakar, and Rao Anwer. Thinking beyond labels: Vocabulary-free fine-grained recognition using reasoning-augmented LMMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16855–16864, 2026.

Ruoyi Du, Dongliang Chang, Kongming Liang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. On-the-fly category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11691–11700, 2023.

Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. XCon: Learning with experts for fine-grained category discovery. In *British Machine Vision Conference (BMVC)*, 2022.

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9284–9292, 2021.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. In *Conference on Language Modeling (COLM)*, 2025.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Pengcheng Guo, Yonghong Song, and Boyu Wang. Towards debiased generalized category discovery. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 5271–5279, 2025. doi: 10.24963/ijcai.2025/587.

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28734–28743, 2026.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8400–8408, 2019.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Proceedings of the International Conference on Learning Representations*, 2020.

Shaozhe Hao, Kai Han, and Kwan-Yee K. Wong. CiPR: An efficient framework with cross-instance positive relations for generalized category discovery. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=1fNcpcdr1o.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *International Conference on Learning Representations (ICLR)*, 2025a.

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

Zhenqi He, Yuanpei Liu, and Kai Han. Category discovery: An open-world perspective. *arXiv preprint arXiv:2509.22542*, 2025b.

Zhenqi He, Yuanpei Liu, and Kai Han. SEAL: Semantic-aware hierarchical learning for generalized category discovery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025c.

Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, et al. GLM-5V-Turbo: Toward a native foundation model for multimodal agents. *arXiv preprint arXiv:2604.26752*, 2026.

Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *International Conference on Learning Representations*, 2018.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642, 2024.

Zhong Ji, Shuo Yang, Jingren Liu, Yanwei Pang, and Chen Tang. A fresh look at generalized category discovery through non-negative matrix factorization. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(6):8274–8288, 2026.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3182–3196, 2024a. doi: 10.18653/v1/2024.findings-emnlp.181.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-V: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024b.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations (ICLR)*, 2025.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in neural information processing systems*, volume 33, pp. 18661–18673, 2020.

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

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Sehyun Kwon, Jaeseung Park, Minkyu Kim, Jaewoong Cho, Ernest K. Ryu, and Kangwook Lee. Image clustering conditioned on text criteria. In *International Conference on Learning Representations*, 2024.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. UME-R1: Exploring reasoning-driven generative multimodal embeddings. In *International Conference on Learning Representations (ICLR)*, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023.

Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27890–27902, 2024.

Haonan Lin, Wenbin An, Jiahao Wang, Yan Chen, Feng Tian, Mengmeng Wang, Guang Dai, Qianying Wang, and Jingdong Wang. Flipped classroom: Aligning teacher attention with student in generalized category discovery. In *Thirty-eighth Conference on Neural Information Processing Systems*, 2024.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. Visual representations inside the language model. In *Conference on Language Modeling (COLM)*, 2025a.

Chunxu Liu, Jiyuan Yang, Ruopeng Gao, Yuhan Zhu, Feng Zhu, Rui Zhao, and Limin Wang. Reasoning guided embeddings: Leveraging MLLM reasoning for improved multimodal retrieval. *arXiv preprint arXiv:2511.16150*, 2025b.

Duo Liu, Zhiquan Tan, Linglan Zhao, Zhongqiang Zhang, Xiangzhong Fang, and Weiran Huang. Generalized category discovery via reciprocal learning and class-wise distribution regularization. In *International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 39651–39668, 2025c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 34892–34916, 2023.

Hezhao Liu, Jiacheng Yang, Junlong Gao, Mengke Li, Yiqun Zhang, Shreyank N Gowda, and Yang Lu. SECOS: Semantic capture for rigorous classification in open-world semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39627–39636, 2026a.

Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *International Conference on Learning Representations*, 2024.

Mingxuan Liu, Zhun Zhong, Jun Li, Gianni Franchi, Subhankar Roy, and Elisa Ricci. Organizing unstructured image collections using natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8907–8918, 2026b.

Yuanpei Liu and Kai Han. DebGCD: Debiased learning with distribution guidance for generalized category discovery. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuanpei Liu, Zhenqi He, and Kai Han. Hyperbolic category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9891–9900, 2025d.

Yuanpei Liu, Zhenqi He, Jialu Tang, and Kai Han. CloSeR: Unified relational distillation from closed-set teachers for category discovery. In *European Conference on Computer Vision (ECCV)*, 2026c.

Shijie Ma, Fei Zhu, Xu-Yao Zhang, and Cheng-Lin Liu. ProtoGCD: Unified and unbiased prototype learning for generalized category discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:6022–6038, 2025.

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

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations*, 2023.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Jona Otholt, Christoph Meinel, and Haojin Yang. Guided cluster aggregation: A hierarchical approach to generalized category discovery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2618–2627, 2024.

Rabah Ouldnoughi, Chia-Wen Kuo, and Zsolt Kira. CLIP-GCD: Simple language guided generalized category discovery. *arXiv preprint arXiv:2305.10420*, 2023.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15691–15701, 2023.

Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yuxun Qu, Minyu Zhou, Yongqiang Tang, Chenyang Zhang, and Wensheng Zhang. Synergistic dual-branch adaptation for multi-modal generalized category discovery. *arXiv preprint arXiv:2606.21446*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Sarah Rastegar, Hazel Doughty, and Cees Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72794–72818, 2023.

Sarah Rastegar, Mohammadreza Salehi, Yuki M. Asano, Hazel Doughty, and Cees G. M. Snoek. SelEx: Self-expertise in fine-grained generalized category discovery. In *European Conference on Computer Vision (ECCV)*, pp. 440–458. Springer, 2024.

Sarah Rastegar, Mina Ghadimi Atigh, Pascal Mettes, Yuki M. Asano, and Cees G. M. Snoek. Fourier self-supervision for fine-grained generalized category discovery. In *Computer Vision – ECCV 2026*, Lecture Notes in Computer Science, pp. 539–557. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37553-7_30.

Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55854–55875, 2025.

Bhupendra Solanki, Ashwin R Nair, Mainak Singha, Souradeep Mukhopadhyay, Ankit Jha, and Biplab Banerjee. GraphVL: Graph-enhanced semantic modeling via vision-language models for generalized class discovery. In *Proceedings of the Fifteenth Indian Conference on Computer Vision Graphics and Image Processing (ICVGIP)*, pp. 1–10. ACM, 2024.

Andreas Stephan, Lukas Miklautz, Kevin Sidak, Jan Philip Wahle, Bela Gipp, Claudia Plant, and Benjamin Roth. Text-guided image clustering. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2960–2976. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.180.

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*, pp. 297–307, 2025.

Luyao Tang, Kunze Huang, Chaoqi Chen, and Cheng Chen. Bures-isotropy alignment: Manifold learning of generalized category discovery. In *The Fourteenth International Conference on Learning Representations*, 2026.

Colin Troisemaine, Vincent Lemaire, Stéphane Gosselin, Alexandre Reiffers-Masson, Joachim Flocon-Cholet, and Sandrine Vaton. Novel class discovery: an introduction and key concepts. *arXiv preprint arXiv:2302.12028*, 2023.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *Proceedings of the International Conference on Learning Representations*, 2022a.

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*, 2022b.

Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19962–19989, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 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, 2025a.

Hongjun Wang, Sagar Vaze, and Kai Han. SPTNet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3QLkwU40EE.

Hongjun Wang, Sagar Vaze, and Kai Han. HiLo: A learning framework for generalized category discovery robust to domain shifts. In *International Conference on Learning Representations (ICLR)*, 2025b.

Menglin Wang, Zhun Zhong, and Xiaojin Gong. Prior-constrained association learning for fine-grained generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21162–21170, 2025c.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025d.

Xinyue Wang, Yuheng Jia, Hui Liu, and Junhui Hou. ESMC: MLLM-based embedding selection for explainable multiple clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26588–26596, 2026.

Ying Wang, Mengye Ren, and Andrew Gordon Wilson. In-context clustering with large language models. *arXiv preprint arXiv:2510.08466*, 2025e.

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

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

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*, 133 (7):4014–4041, 2025. doi: 10.1007/s11263-024-02343-w.

Xingyu Yang, Yu Zhang, Siya Mi, and Xiu-Shen Wei. TAR: Token-aware refinement for fine-grained generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31995–32004, 2026.

Bo Ye, Kai Gan, Tong Wei, and Min-Ling Zhang. Sparsity hurts: Simple linear adapter can boost generalized category discovery. *arXiv preprint arXiv:2605.08183*, 2026.

Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. PromptCAL: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3479–3488, 2023a.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. ClusterLLM: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13903–13920. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.858.

Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems*, 34:22982–22994, 2021.

Bingchen Zhao and Kai Han. Generalized category discovery under the long-tailed distribution. *arXiv preprint arXiv:2506.12515*, 2025.

Bingchen Zhao and Oisin Mac Aodha. Incremental generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised Gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16623–16633, 2023.

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

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*, pp. 3563–3573, 2026a.

Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Generalized fine-grained category discovery with multi-granularity conceptual experts. *International Journal of Computer Vision*, 134(8), 2026b.

Wenliang Zhong, Rob Barton, Lucas Goncalves, Kushal Kumar, Feng Jiang, Hehuan Ma, Yuzhi Guo, Vidit Bansal, Karim Bouyarmane, and Junzhou Huang. Universal guideline-driven image clustering via a hybrid LLM agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33749–33760, 2026.

Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10867–10875, 2021.

Jie Zhu, Yiyang Su, and Xiaoming Liu. Can textual reasoning improve the performance of MLLMs on fine-grained visual classification? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9401–9411, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Further analyses

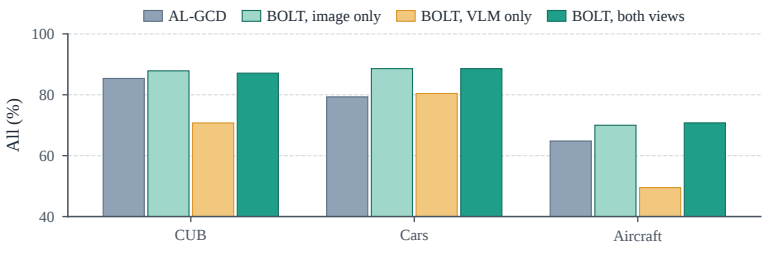


Figure 5: **Missing modality.** Clustering the image feature alone, the VLM projection alone, or both, after training with BOLT, against AL-GCD (the pairs of Table 1).

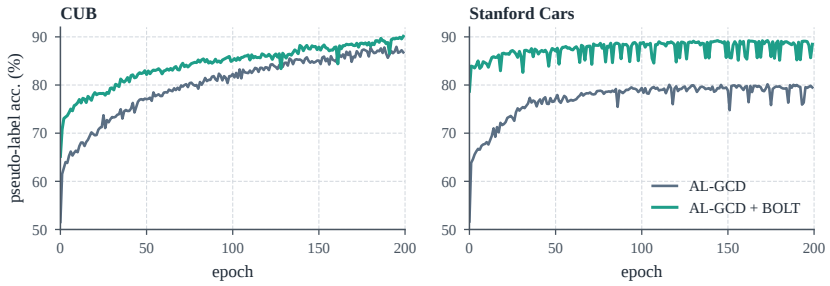


Figure 6: **Pseudo-label accuracy** of the per-epoch clustering on the unlabelled images of CUB and Stanford Cars, AL-GCD and AL-GCD + BOLT (the runs of Table 1). With BOLT the pseudo-labels are more accurate from the first epoch, and the gap persists.

Missing modality and pseudo-label quality. Each view is trained to the same partition, so each keeps it (Figure 5). With the VLM removed at test time, the image model alone still beats the reproduced baseline on every dataset where BOLT gains, and on Stanford Cars it keeps the whole gain (88.6 against 79.3): the knowledge has moved into the vision encoder through the averaged view. With the vision encoder removed, the VLM projection alone still clusters, at 50–80%, lowest on FGVC-Aircraft, where the VLM is weakest. The pseudo-labels the baseline trains on are more accurate from the first epoch and the gap persists (Figure 6), which is the route by which the vote reaches every term of the baseline.

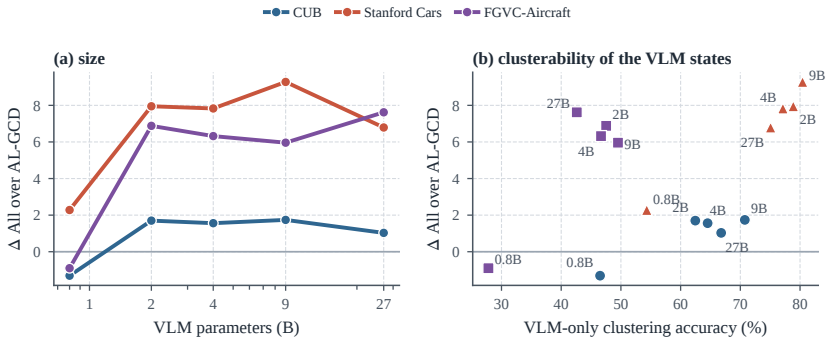


Figure 7: **Gain against the VLM** (Qwen3.5, 0.8B–27B). (a) Gain over AL-GCD against VLM parameters: not monotone (27B $<$ 9B). (b) The same runs against the clustering accuracy of the VLM projection alone at the end of training: the gain appears once the VLM view clusters well enough.

VLM size. We replaced Qwen3.5-9B by four other sizes of the same family, from 0.8B to 27B parameters, keeping everything else fixed (Figure 7; two Gemma-4 models (Gemma Team, 2026) are in Table 6). The gain needs a VLM of roughly 2B parameters and is not improved by 27B. It is not explained by naming either: Qwen3.5-27B names CUB species more accurately than 9B (32.6% vs 25.2% exact matches) yet gains less. What predicts the gain is how well the VLM's states cluster the data by themselves, relative to the baseline: on CUB and Cars the gain is positive above about 60% and small or negative below it, with the loss falling mostly on the novel classes, and on FGVC-Aircraft, where the baseline itself is at 65%, every view that clusters at 43–50% gains 6.0–7.6 points while the weakest (0.8B, 28%) loses 0.9. This clusterability is measurable before training.

When does the VLM help?

The baseline. Figure 8 shows the CLIP and DINOv1 baselines with and without BOLT next to the accuracy of the VLM view alone; Tables 1–2 give the DINOv2 and DINOv3 baselines. On FGVC-Aircraft every baseline gains, and on Stanford Cars every baseline but one, by up to 31 points; the exception, Hyp-SelEx on DINOv2, already clusters Cars at 83.6, the level BOLT brings the other

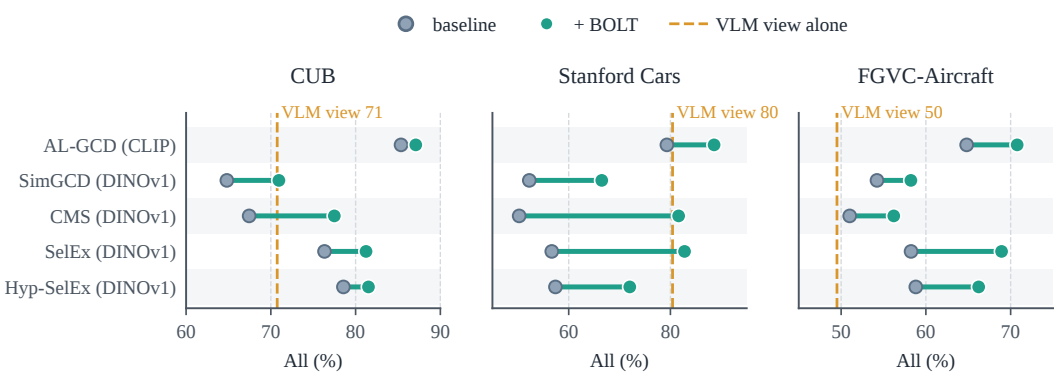


Figure 8: **The CLIP and DINOv1 baselines with and without BOLT**, All accuracy (Table 1); dashed: the VLM view clustered on its own. The gain is large for baselines left of the line and small far right of it.

DINO baselines to, and does not move (-0.8). On CUB the gain broadly falls as the baseline gets stronger: +6.1 and +10.1 for SimGCD and CMS (64.8 and 67.4), +4.9 and +3.0 for SelEx and Hyp-SelEx on DINOv1 (76.3 and 78.6), +1.7 for AL-GCD (85.4), and none for the DINOv2 and DINOv3 baselines (84.1–90.2), which cluster CUB far above the VLM view (70.8). Known classes gain for every CUB baseline (+0.3 to +8.8), while novel classes stop gaining once the baseline is well above the VLM view, the regime in which a noisier view cannot overturn the baseline's wrong partitions without also overturning right ones. Accuracy alone is not the criterion: on FGVC-Aircraft the VLM view clusters at only 49.5% yet every baseline gains, even above 80%, because the condition is on the overlap of the two views' errors, not on the VLM's accuracy: a language model's confusions among aircraft variants differ from a vision model's. Measured on CUB and Cars (Appendix G), the two views' errors are correlated but far from redundant – the VLM view is right on 26–34% of the images the image view gets wrong. Appendix K shows the other side: the CUB classes that lose most are look-alike species to which the VLM gives the same name.

B Algorithm

Algorithm 1 gives BOLT added to a GCD baseline (Section 3).

C Adding BOLT to each baseline

Every baseline receives the term of Equation 2 with its own supervised weight and level weights, and the pseudo-label levels and the VLM vote are built as in Section 3: semi-supervised k-means on $[\hat{f}; \hat{e}]$ with the labelled images pinned, the numbers of known and novel clusters halved from one level to the next. Each baseline keeps its released hyperparameters, and each baseline and its version with BOLT run on the same hardware. **AL-GCD** (CLIP ViT-B/16): the analogical text branch is unchanged; BOLT keeps the baseline's default $\lambda = 0.35$ ($\lambda = 0.5$ in Appendix D); the baseline's agglomerative evaluation runs on $[\hat{f}; \hat{e}]$. **SelEx** (DINOv1 ViT-B/16, DINOv2 ViT-B/14 with registers, DINOv3 ViT-B/16 with the CLS token): its own levels are used; 100 epochs, fine-tuning from block 10 (block 9 on Cars); semi-supervised k-means evaluation on $[\hat{f}; \hat{e}]$. **SimGCD** (DINOv1): levels built each epoch as above, the tail of its 256-dimensional projector supervised, the term ramped in over the first 30 epochs; evaluation by its own classifier head. **CMS** (DINOv1): the tail of its 768-dimensional embedding is supervised and the term added to its supervised loss; evaluation by its own mean-shift and agglomerative clustering at the true number of classes, on its own embedding. The released evaluation script does not run as published (an argument declared twice and an import of a missing function) and the released training script never uses its seed; we fixed both in both arms. **Hyp-SelEx** (DINOv1 and DINOv2): the tail of its 8192-dimensional Poincaré-ball projection is supervised, with h rescaled to the tail's norm so that it stays inside the

Algorithm 1: BOLT added to a GCD baseline

Input : labelled images $\mathcal{D}_l$ with classes, unlabelled images $\mathcal{D}_u$; the baseline's encoder f_θ, projection g, loss $\mathcal{L}_{\text{base}}$ (Eq. 1) and supervised weight λ; a frozen VLM and a taxonomy prompt with keys $1, \dots, K$; number of levels L

Output : the trained encoder; test-time clustering on $c = \frac{1}{\sqrt{2}}[\hat{f}; \hat{e}]$, or on $\hat{f}$ alone

▷ *once per image, before training (no labels used)*

for $x \in \mathcal{D}_l \cup \mathcal{D}_u$ **do**
  ask the VLM for the taxonomy of x; re-encode image, prompt and answer in one teacher-forced pass
  $s_k(x) \leftarrow$ last-layer state at the last token of the value of key k, for $k = 1, \dots, K$

for $k = 1, \dots, K$ **do** remove the mean and top eight principal directions of $\{s_k(x)\}_x$, normalise

$s(x) \leftarrow \mathrm{PCA}_{512}\big([s_1(x); \dots; s_K(x)]\big)$

initialise the projection $P(s) = W\, e(s)$ with bottleneck $e(s) = \mathrm{normalize}(\mathrm{MLP}(s))$

for *each epoch* **do**
  ▷ *the VLM votes in the clustering*
  $c(x) \leftarrow \frac{1}{\sqrt{2}}\,[\hat{f}_\theta(x);\ e(s(x))]$ for every image
  $y^{(0)} \leftarrow$ semi-supervised k-means on c with the labelled images pinned to their classes
  $y^{(i)} \leftarrow$ level i of the baseline's hierarchy, halving the known and novel clusters, $i = 1, \dots, L-1$
  for *each minibatch B (two augmented views per image)* **do**
    $z \leftarrow g(f_\theta(x)) = [z_1; z_2]; \quad h \leftarrow P(s(x)) \cdot \|z_2\| / \|P(s(x))\|$ (no gradient through the scale)
    ▷ *three views of the spare half, nested prefixes $m_i = D/2^{i+1}$*
    $\mathcal{L}_{\text{BOLT}} \leftarrow \sum_{i=0}^{L-1} 2^{-(i+1)} \frac{1}{3} \sum_{u \in \{z_2,\, h,\, \frac{1}{2}(z_2+h)\}} \mathcal{S}\big(u^{:m_i}, y_B^{(i)}\big)$
    update θ, g and P on $\mathcal{L}_{\text{base}}(z, y) + \frac{\lambda}{2}\,\mathcal{L}_{\text{BOLT}}$

ball, and the term is added to both of its supervised losses (hyperbolic distance and angle); 200 epochs with the released curvature and clipping radius; semi-supervised k-means evaluation on $[\hat{f}; \hat{e}]$.

D BOLT AT A LARGER SUPERVISED WEIGHT

Every table in the paper uses the baseline's $\lambda = 0.35$ for both arms. Table 4 gives AL-GCD + BOLT at $\lambda = 0.5$, the best value of the sweep in Figure 4b, against AL-GCD at its default 0.35 (a larger weight lowers the baseline itself, Section 4.3).

Table 4: **BOLT at** $\lambda = 0.5$ against AL-GCD at its default $\lambda = 0.35$, All / Known / Novel (%). The baseline rows are separate runs of the configuration of Table 1, on the same hardware as their BOLT row.

	AL-GCD			**AL-GCD + BOLT**			**Δ**		
Dataset	All	Known	Novel	All	Known	Novel	All	Known	Novel
CUB	85.4	76.7	89.7	87.9	82.2	90.8	+2.6	+5.5	+1.1
Stanford Cars	79.3	92.2	73.1	88.7	93.5	86.4	+9.4	+1.2	+13.3
FGVC-Aircraft	63.1	66.7	61.3	72.6	68.8	74.5	+9.5	+2.2	+13.2
Mean (3 datasets)	75.9	78.6	74.7	83.1	81.5	83.9	+7.2	+3.0	+9.2

Table 5: **AL-GCD and AL-GCD + BOLT at the same λ** on CUB, All accuracy (%) and the gain of BOLT (All / Known / Novel). $\lambda = 0$ removes every supervised term.

λ	+ BOLT	AL-GCD	Δ All / Known / Novel
0	53.8	–	–
0.2	85.4	81.8	+3.5 / +3.1 / +3.8
0.35	87.1	85.4	+1.7 / +5.3 / −0.0
0.5	87.9	83.3	+4.6 / +6.5 / +3.6
0.65	87.3	83.7	+3.6 / +4.8 / +3.0
0.8	86.6	83.6	+3.0 / +4.8 / +2.1
1	86.3	83.0	+3.4 / +7.7 / +1.2

E VLM SIZE IN DETAIL

Table 6 lists the runs of Figure 7: seven VLMs in place of Qwen3.5-9B, everything else fixed, with the gain over AL-GCD on CUB, Stanford Cars and FGVC-Aircraft and the accuracy of the VLM projection clustered alone.

Table 6: **VLM size**, Δ over AL-GCD ($\lambda = 0.35$ for both) and the clustering accuracy of the VLM projection alone after training. A dash: not run.

	CUB				**Stanford Cars**				**FGVC-Aircraft**			
VLM	Δ All	Δ Known	Δ Novel	VLM-only	Δ All	Δ Known	Δ Novel	VLM-only	Δ All	Δ Known	Δ Novel	VLM-only
Gemma-4-E2B	−2.1	+0.8	−3.5	44.2	−0.3	−0.7	−0.1	48.3	+2.8	+2.2	+3.1	30.2
Qwen3.5-0.8B	−1.3	+1.1	−2.5	46.5	+2.3	+0.5	+3.1	54.3	−0.9	+0.8	−1.7	27.8
Qwen3.5-2B	+1.7	+3.0	+1.0	62.5	+8.0	+1.8	+10.9	78.9	+6.9	+4.3	+8.2	47.5
Qwen3.5-4B	+1.6	+5.1	−0.2	64.5	+7.8	+1.4	+10.9	77.1	+6.3	+2.6	+8.2	46.7
Qwen3.5-9B	+1.7	+5.3	0.0	70.8	+9.3	+1.9	+12.8	80.4	+6.0	+4.3	+6.8	49.5
Qwen3.5-27B	+1.0	+3.7	−0.3	66.8	+6.8	+1.6	+9.3	75.1	+7.6	+6.7	+8.1	42.6
Gemma-4-31B	−3.5	+4.0	−7.2	53.3	+7.0	+1.8	+9.5	76.9	+2.5	+5.2	+1.2	51.0

F PROTOCOL AND MEASUREMENT NOISE

Run-to-run variation on these benchmarks is large enough to swallow many published effect sizes. Six runs of one configuration on CUB span 2.2 points; two identical commands in the same job array differ by 1.1; identical runs on two hardware types differ by up to 5.4 points on CUB. We therefore compare every method with its baseline reproduced on the same hardware and report the difference, and we report the model at the end of training rather than the best epoch. AL-GCD's headline clustering is agglomerative (Ward); Table 7 recomputes the CUB and Cars comparisons with semi-supervised k-means, the convention of SelEx and GCD, from the same runs. Per run the two differ by at most 1.1 points, and both CUB gains are larger under k-means.

Table 7: **Ward vs semi-supervised k-means**, same runs, Δ over AL-GCD under each metric.

		Δ Ward			**Δ SS-k-means**		
Dataset	variant	All	Known	Novel	All	Known	Novel
CUB	BOLT	+1.7	+5.3	0.0	+2.8	+4.4	+2.0
CUB	state before value	+2.8	+4.2	+2.1	+3.1	+4.4	+2.5
Stanford Cars	BOLT	+9.3	+1.9	+12.8	+9.1	+1.9	+12.6
Stanford Cars	state before value	+8.3	+2.0	+11.3	+7.9	+1.9	+10.7

G HOW MUCH DO THE TWO VIEWS' ERRORS OVERLAP?

The vote can only correct the image view where the VLM view is right, so what matters is how much the two views' errors overlap. We measure it on the unlabelled images of CUB and Stanford Cars with the final AL-GCD + BOLT model: the image view f and the VLM representation s

(the de-coned per-level states that enter the projection P; the trained bottleneck is not stored in the checkpoints, so this is the VLM view before training) are each clustered with the baseline's semi-supervised k-means and matched to the classes, and an image counts as an error of a view when that view assigns it to the wrong class. Table 8 reports the error rate of each view, of both together, the rate both would share if their errors were independent, the fraction of the image view's errors on which the VLM view is right, and the ϕ correlation of the two error indicators. The errors are positively correlated (ϕ 0.25 on CUB, 0.35 on Cars) but far from redundant: the VLM view is right on 26–34% of the images the image view gets wrong, on known and novel classes alike, which is the room the vote has to correct the clustering. The correlation is higher on Cars than on CUB: there the VLM view is more accurate and its remaining errors sit on the images that are hard for both views. The same holds for pairs, which is what the clustering acts on: of the same-class pairs, the image view splits 16.0% on CUB and 14.5% on Cars, the VLM view 48.5% and 29.2%, and both 12.7% and 10.1% – above the 7.8% and 4.2% of independent errors, below either view alone.

Table 8: **Error overlap of the image view and the VLM view**, unlabelled training images, AL-GCD + BOLT, each view clustered alone with semi-supervised k-means and matched to the classes. *both*: images both views get wrong; *if independent*: the product of the two error rates; ϕ: correlation of the two error indicators.

		error rate (%)					
Dataset	classes	image f	VLM s	both	if independent	s right on f errors (%)	ϕ
CUB	All	12.7	40.9	9.3	5.2	26.7	0.25
	Known	17.9	38.7	12.9	6.9	27.6	0.32
	Novel	10.1	41.9	7.5	4.3	26.0	0.22
Stanford Cars	All	11.3	24.0	7.5	2.7	33.7	0.35
	Known	5.8	17.7	3.9	1.0	31.9	0.33
	Novel	14.0	27.0	9.2	3.8	34.1	0.35

H COST

BOLT costs one forward pass of the frozen VLM per training image, computed once (0.24–0.68 s per image on a single GPU; Table 9) and shared by every run on that dataset, and a 10M-parameter projection. Training is 6–44% slower, the difference being the per-epoch semi-supervised k-means on $[\hat{f}; \hat{e}]$ and the three extra supervised-contrastive terms. Nothing is added at test time: the image feature alone keeps the gain (Section A).

Table 9: **Cost of BOLT.** One VLM pass per training image (Qwen3.5-9B, 448 px), computed once and reused by every run; training wall-clock of AL-GCD and AL-GCD + BOLT for 200 epochs on one GPU (same box, up to eight concurrent jobs, so times are approximate), with seconds per epoch for baseline / + BOLT. BOLT adds 9.96M trainable parameters and nothing at test time when the VLM is dropped (Section A).

		VLM pass (once)		Training, 200 epochs			
Dataset	images	minutes	s / image	AL-GCD	+ BOLT	s / epoch	overhead
CUB	5 994	24	0.24	2.0 h	2.9 h	37 / 53	+44%
Stanford Cars	8 144	86	0.63	2.9 h	3.7 h	52 / 66	+27%
FGVC-Aircraft	6 667	76	0.68	2.7 h	2.9 h	49 / 52	+6%

I THE LEVEL READING OF THE BASELINES' RELEASED CODE

SelEx builds its pseudo-label hierarchy by fitting each level $i \geq 1$ with semi-supervised k-means on the images assigned to known clusters (pinned) and the rest, and the released implementation – inherited by AL-GCD – returns that level's labels in the concatenated order [known-assigned ; rest] while the training loop reads them in data-loader order. For unlabelled images, level $i \geq 1$ therefore

holds another image's label. Level 0, which is fitted on [labelled ; unlabelled] and read in the same order, is unaffected, and so is the evaluation.

The consequence depends on how the dataset is stored. On datasets stored sorted by class, neighbouring images in the loader share a class (CUB 96.7%, FGVC-Aircraft 97.0%), and so do the images whose labels they receive: the misassigned labels encode the file order and, through it, the true class. On CUB, the level-1 labels of unlabelled images have pair precision 0.286 as read by the released code, 0.374 when read correctly, and 0.005 at chance, although the two readings agree on only 1.7% of the images. On datasets not stored by class (Stanford Cars 0.6%) the misassigned labels are at chance: noise, not information.

The published numbers of SelEx and AL-GCD were obtained with the released code (and so, to the extent that they build on it, were those of later SelEx-based methods), and so were all results in the main text: the baseline and the baseline with BOLT always share the released reading, so the differences compare like with like and the absolute numbers follow the published protocol. Our SimGCD port has no such level code. Table 10 repeats the FGVC-Aircraft pair with a corrected reading (a one-line change that returns the levels in loader order): the baseline changes by -9.7 points All, consistent with the file order carrying class information, and BOLT changes All / Known / Novel by +4.0 / +12.7 / -0.3 under the corrected reading against +9.5 / +2.2 / +13.2 under the released one.

Table 10: **The same pair under the released and a corrected reading** of the pseudo-label levels, All / Known / Novel (%), AL-GCD (CLIP ViT-B/16), BOLT at $\lambda = 0.5$ (Appendix D), same hardware and settings within a row.

	AL-GCD			AL-GCD + BOLT			Δ		
Dataset (level reading)	All	Known	Novel	All	Known	Novel	All	Known	Novel
FGVC-Aircraft (released)	63.1	66.7	61.3	72.6	68.8	74.5	+9.5	+2.2	+13.2
FGVC-Aircraft (corrected)	53.3	57.7	51.1	57.4	70.5	50.8	+4.0	+12.7	−0.3

J PROMPTS AND THE LEAK CHECK

Each image, letterboxed to 448 pixels, is answered once with a JSON taxonomy with fixed keys from the broadest category down to one level finer than the dataset's labels (CUB: `domain, class, group, family, genus, species, subtype`; Cars: `domain, kind, body, make, model, generation, trim`; Aircraft: `domain, kind, role, manufacturer, family, model, variant`). The model writes `unknown` for a level it cannot support and for every finer level. Thinking mode is off. The prompts show the format with an example from an unrelated domain and give no list of classes: listing the known classes, even "as a guide", makes the model file novel images under known names.

Leak check. Every example phrase in a prompt is compared with every class name, known and novel, after normalisation. No prompt names a novel class. The CUB prompt contains one known class, which GCD permits since known class names are given: it uses "Black-footed Albatross (not Phoebastria nigripes)" to ask for common rather than Latin names.

The prompts, verbatim. The image and the text below form a single user turn to Qwen3.5-9B, and the answer is decoded greedily. The CUB prompt is shown in full, as the model receives it (long lines are wrapped for display only). The Stanford Cars and FGVC-Aircraft prompts are the same text with four parts replaced – the first sentence, the key line, the level definitions and the evidence rule – shown in the two boxes after it; every other line is identical. Every answer on the three datasets parsed as JSON (5 994, 8 144 and 6 667 training images); the label-level key (species, model, model) is `unknown` for 6.8%, 1.3% and 9.5% of the images.

Prompt for CUB

```
<image: the training image, letterboxed to 448 x 448 pixels>
Identify what this image shows as a strict taxonomic hierarchy, running from the
broadest level down to a level finer than an ordinary label.

Reply with ONE JSON object on a single line and nothing else. Use exactly these
seven keys, in this order:

{"domain": "", "class": "", "group": "", "family": "", "genus": "", "species": "",
"subtype": ""}

What each level means, coarse to fine:
- domain   the broadest kind of thing         e.g. animal, plant, vehicle
- class    its major category                 e.g. bird, mammal, car
- group    a broad functional grouping        e.g. seabird, songbird, bird of prey
- family   the recognisable family            e.g. albatross, warbler, gull
- genus    a finer grouping inside the family e.g. Phoebastria, Setophaga
- species  the specific type, as its COMMON English name, not a Latin binomial
                                              e.g. Black-footed Albatross (not
                                              Phoebastria nigripes)
- subtype  finer than species                 e.g. adult male, dark morph, juvenile,
                                                   breeding plumage

Rules:
- Every key must be present, with a non-empty string value.
- Each level must be a subtype of the level above it.
- Go as deep as the visible evidence supports. If the image does not support a
  level, write "unknown" for that level AND for every finer level below it. A
  confident coarse answer beats a guessed fine one.
- Judge from visible evidence only: shape, proportions, colour and pattern of each
  part, posture, setting.
- There is no fixed list of allowed answers. Name what you actually see.
- Output the JSON object only. No preamble, no explanation, no markdown fence.
```

Stanford Cars: the lines that differ from the CUB prompt

```
Identify the vehicle in this image as a strict hierarchy, running from the broadest
level down to a level finer than an ordinary label.
{"domain": "", "kind": "", "body": "", "make": "", "model": "", "generation": "",
"trim": ""}
- domain      the broadest kind of thing              e.g. vehicle, animal, building
- kind        the major category of vehicle           e.g. car, pickup truck, van,
                                                           motorcycle, bus
- body        the body style                          e.g. sedan, SUV, coupe,
                                                           convertible, hatchback,
                                                           wagon, minivan
- make        the manufacturer                        e.g. Peugeot, Skoda, Renault,
                                                           Opel
- model       the model name, without the year        e.g. 308, Octavia, Clio, Astra
- generation  the model years or generation           e.g. 2014-2020, or a
                                                           generation label such as
                                                           Mk3
- trim        the trim or variant                     e.g. GT Line, vRS, Initiale
                                                           Paris
- Judge from visible evidence only: badges, grille, headlights, body proportions,
  roofline, wheels, windows, tail-lights.
```

FGVC-Aircraft: the lines that differ from the CUB prompt

```
Identify the aircraft in this image as a strict hierarchy, running from the broadest
level down to a level finer than an ordinary label.
{"domain": "", "kind": "", "role": "", "manufacturer": "", "family": "", "model":
"", "variant": ""}
- domain        the broadest kind of thing            e.g. vehicle, animal, building
- kind          the major category of aircraft        e.g. airplane, helicopter,
                                                           glider, airship
- role          what the aircraft is for              e.g. narrow-body airliner,
                                                           wide-body airliner,
                                                           regional jet, turboprop
                                                           airliner, business jet,
                                                           light aircraft, military
                                                           fighter, military
                                                           transport, bomber,
                                                           trainer
- manufacturer  who built it                          e.g. COMAC, Sukhoi, Pilatus,
                                                           Mitsubishi, Honda
                                                           Aircraft, Irkut
- family        the model family                      e.g. 787, A350, C919, Superjet
                                                           100, PC-12, HondaJet
- model         the specific model or series          e.g. 787-9, A350-900,
                                                           C919-100, Superjet
                                                           100-95, A220-300, PC-12
                                                           NGX
- variant       finer than the model                  e.g. 787-9 with raked
                                                           wingtips, A350F
                                                           freighter, A220-300 long
                                                           range, military KC-46
                                                           tanker
- Judge from visible evidence only: fuselage length and cross-section, number and
  placement of engines, wing shape and winglets, tail configuration, nose and
  cockpit windows, landing gear, door count, livery is NOT evidence of type.
```

How an answer is read. The generated answer is parsed, rewritten as one canonical JSON line with the keys in the prompt's order, and appended to the image and the prompt; a single teacher-forced forward pass then gives the last-layer hidden state at the last token of each value, one state per level (Section 3). A level answered `unknown` is read like any other, and since every level below an `unknown` is `unknown` too, each image yields exactly seven states whatever the depth of its answer. The words themselves are never used as labels.

Example answers. Figure 9 shows, for each dataset, three unlabelled training images drawn at random (seed 0) under fixed rules: a known-class image, a novel-class image, and an image whose label-level answer names a wrong class while a coarser level is right. The VLM names the Nighthawk, the Acura RL and the (novel) Ferrari 458 Italia correctly, calls a Northern Waterthrush a Swainson's Thrush and a Falcon 900 a Global 5000, and stops at the family for a Eurofighter Typhoon. For the Lincoln Sparrow, the Audi S6 and the A330-300 it is right down to the genus, the make and the family and wrong only at the finest level (song sparrow, A4, A330-200), between close look-alikes – which is why BOLT reads the states of these answers and never uses their words as labels.

CUB



known-class image

true class:
Nighthawk
(known)

domain	animal
class	bird
group	nocturnal bird
family	nightjar
genus	Caprimulgus
species	Common Nighthawk
subtype	adult in flight



novel-class image

true class:
Northern
Waterthrush
(novel)

domain	animal
class	bird
group	songbird
family	Parulidae
genus	Catharus
species	Swainson's Thrush
subtype	adult



fine wrong, coarse right

true class:
Lincoln Sparrow
(known)

domain	animal
class	bird
group	songbird
family	sparrow
genus	Melospiza
species	song sparrow
subtype	adult male

Stanford Cars



known-class image

true class:
Acura RL Sedan
2012
(known)

domain	vehicle
kind	car
body	sedan
make	Acura
model	RL
generation	2006-2012
trim	unknown



novel-class image

true class:
Ferrari 458
Italia Coupe 2012
(novel)

domain	vehicle
kind	car
body	coupe
make	Ferrari
model	458 Italia
generation	2009-2015
trim	unknown



fine wrong, coarse right

true class:
Audi S6 Sedan
2011
(novel)

domain	vehicle
kind	car
body	sedan
make	Audi
model	A4
generation	B7
trim	unknown

FGVC-Aircraft



known-class image

true class:
Eurofighter
Typhoon
(known)

domain	vehicle
kind	airplane
role	military fighter
manufacturer	BAE Systems
family	Typhoon
model	unknown
variant	unknown



novel-class image

true class:
Falcon 900
(novel)

domain	vehicle
kind	airplane
role	business jet
manufacturer	Bombardier
family	Global Express
model	Global 5000
variant	unknown



fine wrong, coarse right

true class:
A330-300
(known)

domain	vehicle
kind	airplane
role	wide-body airliner
manufacturer	Airbus
family	A330
model	A330-200
variant	unknown

Figure 9: **Example answers** of Qwen3.5-9B to the prompts above, for unlabelled training images drawn at random (seed 0) under fixed rules: a known-class image, a novel-class image, and one whose label-level answer is wrong while a coarser level is right (judged by string match to the class name). Each image is shown in full, letterboxed. The label-level value (CUB `species`, Cars `model`, Aircraft `model`) is green when it names the true class, red when it names another one; `unknown` in grey.

K QUALITATIVE ANALYSIS



Figure 10: **Success and failure cases on novel classes.** Left: unlabelled images of novel classes that AL-GCD + BOLT assigns to their class and AL-GCD does not; right: the reverse (CUB, top; Stanford Cars, bottom). Below each image: the ground-truth class (GT) and the class each model assigns it to (image feature, each model with its own Hungarian matching), green when right and red when wrong. Classes are ranked by the change in their accuracy and one image per class is drawn at random (seed 0); images are shown whole.

Success and failure cases. Figure 10 shows unlabelled images of novel classes that only one of the two models assigns to their class. Where BOLT succeeds, the VLM's own answer for the image is right or close (*indigo bunting*, *gull* for the California Gull, *GMC Yukon*, *Chevrolet Monte Carlo*, *BMW 5 Series* for an M5), and BOLT moves the image out of the look-alike class AL-GCD put it in. Where it fails, the VLM's answer points at the look-alike – a Prairie Warbler called a *yellow warbler* and an Aston Martin Virage a *V8 Vantage*, the very classes BOLT then assigns them to – or misses the distinction the class needs (*Dodge Ram 1500* for a Dakota Crew Cab that BOLT assigns to the Club Cab), and the second view pulls the image toward the look-alike.

Features and attention. Figure 11 and Table 11 use the image feature f of the final AL-GCD and AL-GCD + BOLT models of Table 1, clustered with semi-supervised k-means (the VLM projection is not needed at test time, Section A). Over all 196 Stanford Cars classes, both models form compact class clusters, and with BOLT more of them are a single colour (Figure 11). On CUB, BOLT improves 68 classes and lowers 44. The largest losses are novel classes absorbed by a look-alike (Table 11); in four of the six, the VLM writes the same species name for both classes – Clay-colored and Brewer's Sparrow are both "Chipping Sparrow", Horned and Eared Grebe both "Black-necked Grebe", Pelagic and Red-faced Cormorant both "double-crested cormorant", Herring and Glaucous-winged Gull both "gull". Where the VLM's error coincides with the image model's, the second view pulls the two classes together.

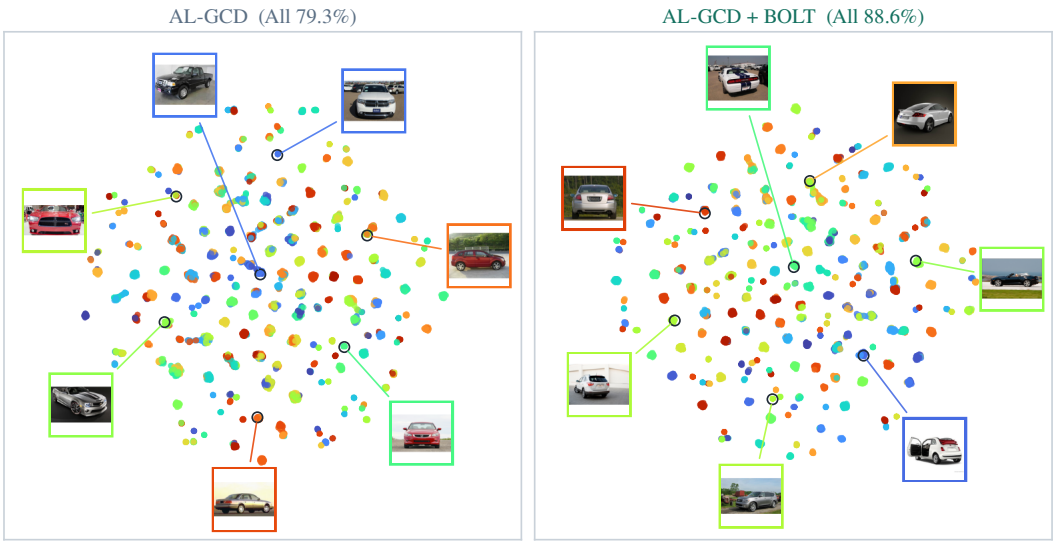


Figure 11: **t-SNE** of the image feature of all unlabelled Stanford Cars images (196 classes), AL-GCD (left) and AL-GCD + BOLT (right), coloured by true class; the titles give each model's All accuracy (Table 1). Thumbnails: for each of seven k-means regions of the map (seed 0), the image nearest the region's centre, shown whole and framed in its class colour.

Table 11: **The six CUB classes on which BOLT loses most** (image feature): accuracy of each run on the class, the class that absorbs it under BOLT, and the species name the VLM writes most often for each.

class	split	AL-GCD	+ BOLT	merged into	VLM species (class)	VLM species (merged into)
Louisiana Waterthrush	novel	90.0	0.0	Northern Waterthrush	Pine Warbler	ovenbird
Prairie Warbler	novel	90.0	0.0	Yellow Warbler	yellow-rumped warbler	yellow warbler
Horned Grebe	known	68.8	18.8	Eared Grebe	Black-necked Grebe	Black-necked Grebe
Clay colored Sparrow	known	35.3	0.0	Brewer Sparrow	Chipping Sparrow	Chipping Sparrow
Pelagic Cormorant	novel	86.7	53.3	Red faced Cormorant	double-crested cormorant	double-crested cormorant
Herring Gull	known	33.3	0.0	Glaucous winged Gull	gull	gull

Figure 12 shows the backbone's attention on the success cases of Figure 10 for SelEx on DINOv1, whose backbone is fine-tuned in its last blocks, with and without BOLT. Both models attend to the object; with BOLT the attention on the GMC Yukon extends from the grille to the wheels and the body outline, while on the birds and the Monte Carlo the two maps are similar.

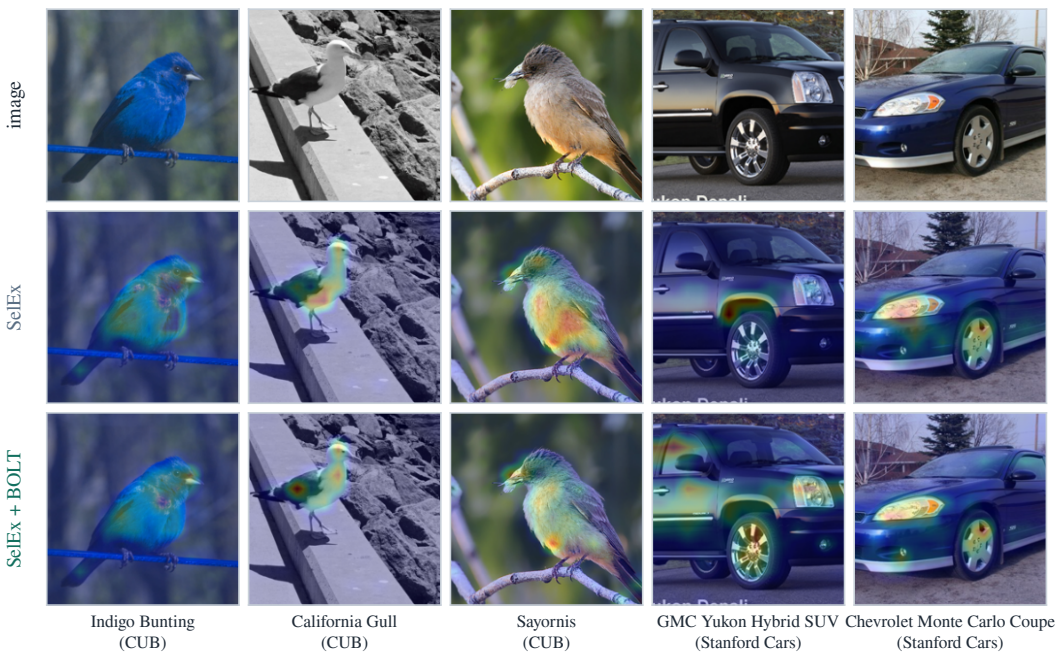


Figure 12: **Attention** of the fine-tuned DINOv1 ViT-B/16 backbone of SelEx and SelEx + BOLT (final epoch) on the success cases of Figure 10: last-layer CLS-to-patch attention averaged over heads.