MULTI-LABEL GENERALIZED CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we introduce Multi-Label Generalized Category Discovery (ML-GCD), a new setting for jointly recognising Known categories and discovering Novel categories in multi-label data. It extends the single-label formulation of Generalized Category Discovery (GCD) to category-wise presence estimation. In ML-GCD, labelled instances contain only Known categories, whereas the unlabelled pool spans Known-only, Novel-only, and mixed instances with all annotations withheld. For unlabelled instances, cross-view category learning derives supervision from model-generated targets, but the resulting consistency objective lacks an independent, fixed reference for representation learning. Our key idea is to couple category discovery with the preservation of global and local visual structure, using a fixed pretrained visual reference independent of evolving category assignments. We propose Patch-guided Discovery (PaD), which transfers Known-category knowledge to unlabelled instances, preserves global feature relations, and retains local visual information. Experiments on VOC2007, MIR-Flickr, and COCO2014 demonstrate the effectiveness of PaD for ML-GCD, with ablation studies supporting the value of fixed visual guidance in complementing model-generated supervision.

1 INTRODUCTION

Generalized Category Discovery (GCD) extends recognition beyond annotated categories by learning from labelled Known examples and unlabelled Known and Novel examples (Vaze et al., 2022). Its single-label formulation assigns one category to each image. Multi-object scenes instead require category-wise predictions: an image containing a Known bicycle and a Novel person requires recognising both, with consistent Novel identities across images.

We introduce Multi-Label Generalized Category Discovery (ML-GCD) for this setting. Labelled images contain only Known categories, whereas the unlabelled pool spans Known-only, Novel-only, and mixed images, with all annotations withheld (Figure 1). The goal is to predict category sets with fixed Known identities and shared anonymous Novel outputs. Section 2 distinguishes this supervision boundary from related multi-label discovery formulations.

Recognising co-occurring Known and Novel categories requires category-specific discrimination within each image. Cross-view category learning constructs targets from model predictions (Wen et al., 2023). When category outputs share an image representation, sparse targets supervise only selected outputs, while the resulting updates affect features shared across categories. Prediction consistency and contrastive learning constrain outputs and feature relations, respectively, but neither explicitly preserves the pretrained relations between images or the local spatial features within them.

Our key idea is to couple category discovery with the preservation of pretrained visual structure. A fixed visual reference guides global relations and local features independently of evolving category targets, without Novel annotations or object-level labels.

We propose Patch-guided Discovery (PaD), which learns sparse cross-view category targets under three complementary constraints from a frozen visual reference. Known-prediction transfer supervises Known outputs on unlabelled images; global relation preservation matches similarities to Known anchors and neighbouring images; aligned patch distillation retains corresponding spatial features across views. These constraints connect available category knowledge to image-level and spatial structure during discovery.

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.

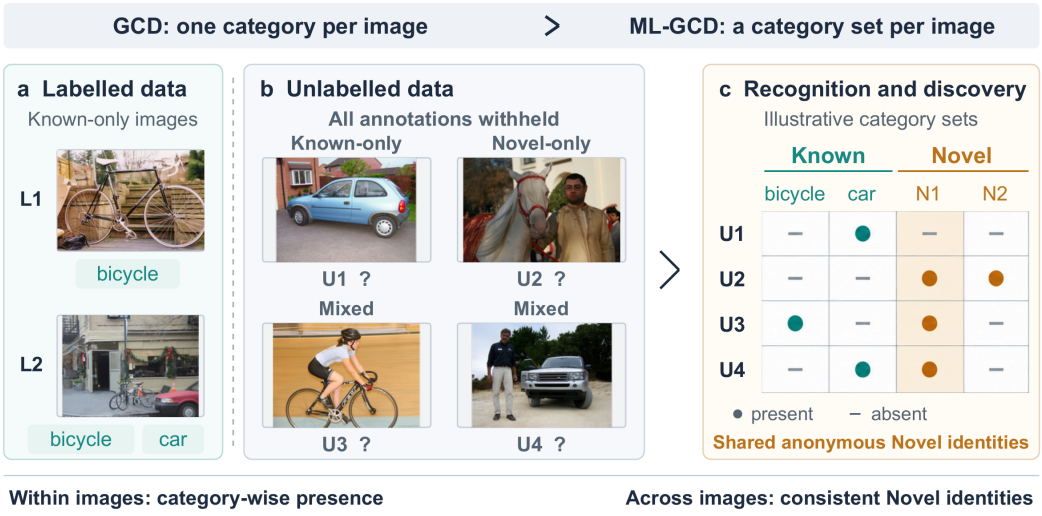


Figure 1: **Multi-Label Generalized Category Discovery.** (a) Labelled Known-only images. (b) Unlabelled Known-only, Novel-only, and mixed images; composition labels are explanatory and all annotations are withheld. (c) Category sets with fixed Known identities and shared anonymous Novel outputs; N1 recurs across U2–U4. Dots/dashes denote presence/absence. VOC2007 examples and target sets illustrate the task, not measured predictions.

PaD improves mean All and Novel mAP and aggregate F1 over the category-learning control on VOC2007, MIRFlickr, and COCO2014. Mixed-image precision improves across datasets, while joint Known–Novel detection improves on MIRFlickr but declines on VOC2007 and COCO2014.

Our contributions are threefold:

1. **We formulate ML-GCD** for joint category recognition and discovery in multi-label images, with an explicit supervision boundary and shared anonymous Novel identities. We establish an evaluation protocol covering category ranking, set prediction, and recognition under different category compositions.

2. **We develop Patch-guided Discovery (PaD)**, coupling evolving category targets with fixed visual guidance. Its prediction, relation, and patch constraints provide complementary supervision by transferring Known-category knowledge and preserving global and local visual structure.

3. **We demonstrate the effectiveness of PaD** across three datasets in category ranking and aggregate set prediction. Component ablations distinguish the roles of prediction transfer and representation preservation, while composition-specific analyses characterise their implications for joint Known–Novel recognition.

2 Related Work

Generalized category discovery. GCD learns Known and Novel categories from labelled and unlabelled data (Vaze et al., 2022). SimGCD studies parametric classification and prediction bias (Wen et al., 2023); PIM uses information maximisation (Chiaroni et al., 2023). PromptCAL learns affinities with visual prompts (Zhang et al., 2023), and SPTNet alternates spatial prompting with model adaptation (Wang et al., 2024). LegoGCD addresses Known-category forgetting (Cao et al., 2024). Recent methods incorporate synthesised text embeddings (GET) (Wang et al., 2025), dominant and contextual visual primitives (ConGCD) (Tang et al., 2025), or token pruning (Attention Focusing) (Xu et al., 2025). These methods assign one category per image; ML-GCD requires predicting multiple co-occurring Known and Novel categories.

Table 1: **Comparison of task formulations.** ✓ denotes an explicit property or requirement; – denotes one not stipulated. MLL-ML is multi-label learning with missing labels. Mixed composition means Known and Novel categories coexist in an example (a bag for DMNL). The last two rows specify supervision restrictions.

Feature	GCD	MLL-ML	DMNL	DLCL	MCUL	**ML-GCD**
Multi-label prediction	–	✓	✓	✓	✓	✓
Novel category discovery	✓	–	✓	✓	✓	✓
Mixed category composition	–	–	✓	✓	✓	✓
Labelled subset excludes Novel	✓	–	–	–	–	✓
Fully unlabelled training pool	✓	–	–	–	–	✓

Multi-label image classification. Multi-label recognition models category dependencies and spatial evidence through label graphs (ML-GCN) (Chen et al., 2019b), semantic-specific representations (SSGRL) (Chen et al., 2019a), and joint visual–label transformers (C-Tran) (Lanchantin et al., 2021). Spatial aggregation uses class-specific attention (Zhu & Wu, 2021), query-based decoding (ML-Decoder) (Ridnik et al., 2023), or state-space modelling (MambaML) (Zhu et al., 2025). Complementary objectives reduce easy-negative contributions with ASL (Ridnik et al., 2021) or exploit multi-label supervised contrastive learning (Zhang & Wu, 2024). ML-GCD further requires learning shared anonymous Novel identities from unlabelled images.

Multi-label discovery and incomplete supervision. Multi-label discovery includes DMNL for bags with Known labels (Zhu et al., 2017), DLCL for latent label columns (Huang et al., 2020), and MCUL for latent columns with missing observed entries (Huang et al., 2021). ML-GCD instead excludes Novel categories from labelled images and withholds all annotations in the unlabelled pool (Table 1). Our experiments adapt representative methods to this boundary.

Category supervision and visual guidance. Balanced cross-view assignments, as used in SwAV (Caron et al., 2020), provide a mechanism for constructing targets without manual labels. PaD uses assignment scores to select sparse supervision for independent sigmoid outputs. It further draws on pretrained DINO features (Caron et al., 2021) and the principle of transferring feature relations (Park et al., 2019). Its local objective matches visible, spatially aligned patches from a frozen encoder, whereas iBOT uses masked prediction with an online tokeniser (Zhou et al., 2022). PaD's fixed reference remains independent of evolving category targets.

3 Multi-Label Generalized Category Discovery

Data and annotation boundary. Let $\mathcal{K}$ and $\mathcal{N}$ be disjoint sets of Known and Novel categories, with $K = |\mathcal{K}|$, $R = |\mathcal{N}|$, and $C = K + R$. The full category vocabulary is $\mathcal{C} = \mathcal{K} \cup \mathcal{N}$. An instance x_i has a binary presence vector $y_i \in \{0,1\}^C$: $y_{ic} = 1$ indicates that category c is present. The annotation-validity mask $v_i \in \{0,1\}^C$ identifies entries eligible for supervision or evaluation. The learner receives

$$\mathcal{D}_L = \{(x_i, y_i^{\mathcal{K}}, v_i^{\mathcal{K}})\}_{i=1}^{n_L}, \qquad \mathcal{D}_U = \{x_i\}_{i=1}^{n_U}, \qquad \|y_i^{\mathcal{K}}\|_0 \geq 1, \quad \|y_i^{\mathcal{N}}\|_0 = 0 \quad (i = 1, \ldots, n_L). \tag{1}$$

Here n_L, n_U are subset sizes; superscripts $\mathcal{K}, \mathcal{N}$ select category groups; and $\|\cdot\|_0$ counts nonzero entries. Indices are local to each subset. $\mathcal{D}_U$ contains Known-only, Novel-only, and mixed instances, with all annotations withheld and positive coverage of every category. Masked entries contribute neither supervision nor evaluation.

Composition-dependent supervision. The task requires transferring Known recognition to category combinations absent from the labelled subset and learning Novel identities from unlabelled images. PaD supervises only Known columns on labelled images; Appendix I studies structural Novel-negative supervision. The supplied Novel category counts are $R = 10$, 12, and 40 for VOC, MIR, and COCO.

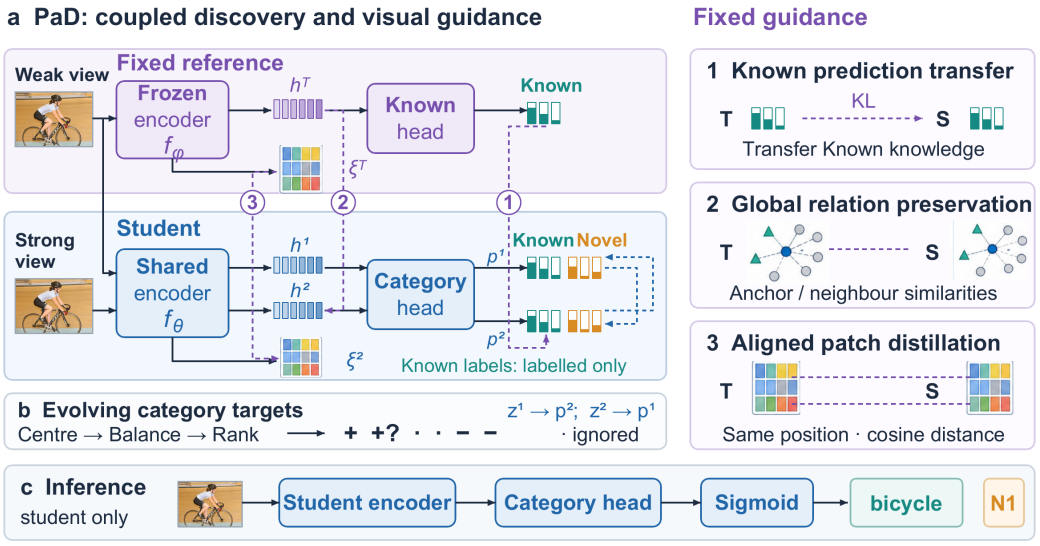


Figure 2: **Patch-guided Discovery (PaD).** (a) A shared student learns from category targets and fixed visual guidance. Superscripts 1/2 denote weak/strong views and T/S teacher/student. Indices are omitted: global/patch features h^T, ξ^T correspond to $\widehat{h}_i, \widehat{\xi}_{it}$, and h^2, ξ^2 to $h_i^{(2)}, \xi_{it}$. Constraints transfer Known predictions (1), anchor/neighbour relations (2), and aligned patches (3) to the strong view. The teacher encoder is frozen; its Known head and anchors are fixed after preparation. Valid Known labels supervise both student views. (b) Detached, centred scores supervise the opposite view through sparse targets. (c) Student-only inference. The VOC photograph is reused across schematic views; features and outputs are illustrative. N1 denotes an anonymous Novel slot.

Prediction and global alignment. Index Known outputs by $c = 1, \ldots, K$ and anonymous Novel outputs by $c = K + r$, where $r = 1, \ldots, R$. Let $p_c(x_i)$ be the predicted presence probability. Inference thresholds each output independently. Novel identities are aligned only for evaluation. Enumerate the true Novel categories by $j = 1, \ldots, R$, and let n_E be the number of evaluation instances. Define

$$A_{rj} = \mathrm{AP}(\{(y_{i,K+j}, p_{K+r}(x_i), v_{i,K+j})\}_{i=1}^{n_E}), \qquad \pi^* = \arg\max_{\pi \in \mathfrak{S}_R} \sum_{r=1}^{R} A_{r,\pi(r)}. \tag{2}$$

Here AP uses valid entries, A is the slot–category AP matrix, and $\mathfrak{S}_R$ contains permutations of R categories. The mapping π^* aligns anonymous Novel outputs for scoring; Known identities remain fixed. All subgroups reuse this mapping. We report All/Known/Novel mAP, macro/micro F1 at threshold 0.5, and cardinality MAE (Appendix C).

4 Patch-guided Discovery

Overview. PaD learns multi-label category assignments from category supervision and fixed visual guidance (Figure 2). Valid Known annotations and sparse cross-view targets supervise category predictions, while a frozen teacher supplies Known-prediction, global-relation, and aligned-patch targets independent of evolving Novel assignments.

4.1 Independent category prediction and Known supervision

The category learner uses independent sigmoid outputs to predict several categories in one image. Let f_θ be the student visual encoder and $h_i^{(a)} \in \mathbb{R}^d$ its ℓ_2-normalised global feature for view $x_i^{(a)}$, where $a \in \{1, 2\}$ indexes the weak and strong views. The trainable student parameters are denoted collectively by θ. With unit-norm classifier direction $w_c \in \mathbb{R}^d$, bias b_c, and unconstrained scale

parameter α_c, the category head computes

$$z_{ic}^{(a)} = s_c \langle w_c, h_i^{(a)} \rangle + b_c, \qquad s_c = \mathrm{softplus}(\alpha_c) + 10^{-4}, \qquad p_{ic}^{(a)} = \sigma(z_{ic}^{(a)}). \tag{3}$$

Here z, p, σ denote logits, probabilities, and the sigmoid function. Category-specific scales and biases calibrate cosine similarities for variable-size category prediction. Appendix A.1 gives dimensions and initialisation.

Known annotations establish the first K output identities. Asymmetric Loss (ASL) reduces easy-negative contributions (Ridnik et al., 2021). For binary target y and logit z, let $p = \sigma(z)$ and $p_m = \max(p - m, 0)$:

$$\ell_{\mathrm{ASL}}(z, y) = -y(1-p)^{\gamma_+} \log p - (1-y) p_m^{\gamma_-} \log(1 - p_m). \tag{4}$$

We use margin $m = 0.05$ and focusing exponents $\gamma_+ = 0$, $\gamma_- = 4$. Let $\mathcal{B}_L, \mathcal{B}_U$ index labelled/unlabelled batch entries, with sizes B_L, B_U and $B = B_L + B_U$. Number instances so that $\mathcal{B}_L \cup \mathcal{B}_U = \{1, \ldots, B\}$. The Known loss is

$$\mathcal{L}_K = \frac{\sum_{a=1}^{2} \sum_{i \in \mathcal{B}_L} \sum_{c=1}^{K} v_{ic} \ell_{\mathrm{ASL}}(z_{ic}^{(a)}, y_{ic})}{2 \max(1, \sum_{i \in \mathcal{B}_L} \sum_{c=1}^{K} v_{ic})}. \tag{5}$$

The mask restricts supervision to valid Known entries in the labelled subset.

4.2 Sparse cross-view category discovery

Sparse cross-view targets provide category supervision for unlabelled images while leaving intermediate-confidence entries unassigned. Let $Z_U^{(a)} \in \mathbb{R}^{B_U \times C}$ collect their logits. Detached logits produce assignment scores

$$Q^{(a)} = \mathrm{Sinkhorn}_3\Big(\exp\left[(\mathrm{sg}(Z_U^{(a)}) - \boldsymbol{\mu})/\tau_{\mathrm{assign}}\right]\Big), \qquad \tau_{\mathrm{assign}} = 0.1. \tag{6}$$

sg stops gradients and $\boldsymbol{\mu} \in \mathbb{R}^C$ is the running logit centre. $\mathrm{Sinkhorn}_3$ alternates category and instance normalisation three times, yielding unit row mass and approximately balanced category mass. Q ranks targets; presence probabilities remain independent sigmoid outputs. Appendix A.3 specifies stabilisation and centre updates.

Each unlabelled instance receives its highest-ranked category as a positive target. The $\mathrm{round}(0.5 B_U)$ instances with the largest second-to-first score ratios receive a second positive. Ranks $q+1, \ldots, C$ are negative; intermediate ranks are ignored. We use $q = 5$ for VOC, with cutoff specifications in Appendix A.1. Let $\widetilde{y}_{ic}^{(a)} \in \{0, 1\}$ be the pseudo-target and $M_{ic}^{(a)} \in \{0, 1\}$ its selection mask. The opposite view is supervised by

$$\mathcal{L}_U = \frac{1}{2} \sum_{(a,b) \in \{(1,2),(2,1)\}} \frac{\sum_{i \in \mathcal{B}_U} \sum_{c=1}^{C} M_{ic}^{(b)} \ell_{\mathrm{ASL}}(z_{ic}^{(a)}, \widetilde{y}_{ic}^{(b)})}{\max(1, \sum_{i \in \mathcal{B}_U} \sum_{c=1}^{C} M_{ic}^{(b)})}. \tag{7}$$

The two directions share student weights; an empty unlabelled batch contributes zero. Training selects approximately 1.5 positive targets per image, whereas inference thresholds all category outputs independently. Fixed teacher targets complement this sparse supervision through prediction, relation, and patch constraints.

4.3 A fixed visual reference

A frozen DINO-v1 ViT-B/16 encoder f_ϕ provides the visual reference (Caron et al., 2021). Both views share a full-image 224×224 resize and horizontal flip; strong view 2 adds photometric perturbations to weak view 1 without changing patch coordinates. The teacher processes view 1, and transfer losses constrain student view-2 features or predictions.

Known-prediction distillation transfers label information to labelled and unlabelled images. Let $h_i = h_i^{(2)}$ and $\widehat{h}_i \in \mathbb{R}^d$ denote normalised student and teacher global features. A linear teacher

head is fitted for 30 epochs by binary cross-entropy on valid labelled Known entries, then frozen to produce probabilities $\widehat{p}_{ic}$ for $c = 1, \ldots, K$. Its distillation loss is

$$\mathcal{L}_{\text{K-KD}} = \frac{1}{BK} \sum_{i=1}^{B} \sum_{c=1}^{K} \mathrm{KL}_{\mathrm{Ber}}(\widehat{p}_{ic} \| p_{ic}^{(2)}), \qquad \mathrm{KL}_{\mathrm{Ber}}(\widehat{p} \| p) = \widehat{p} \log \frac{\widehat{p}}{p} + (1 - \widehat{p}) \log \frac{1 - \widehat{p}}{1 - p}. \tag{8}$$

Here $\mathrm{KL}_{\mathrm{Ber}}$ is Bernoulli KL divergence at temperature one; the frozen predictions guide all batch images.

4.4 Global relation preservation

Global relation preservation transfers the teacher's feature geometry through category anchors and image neighbours. Known anchors first provide reference directions tied to the available annotations. For Known category c, let $\mathcal{W}_c$ be the multiset of valid positive labelled occurrences sampled during teacher preparation. Repeated visits count separately. The frozen anchor is

$$\kappa_c = \mathrm{norm}\left(\sum_{o \in \mathcal{W}_c} \widehat{h}_o \right), \qquad \mathcal{L}_{\text{anchor}} = \frac{1}{B|\mathcal{K}_+|} \sum_{i=1}^{B} \sum_{c \in \mathcal{K}_+} \psi\Big(\langle h_i, \kappa_c \rangle - \langle \widehat{h}_i, \kappa_c \rangle \Big). \tag{9}$$

norm denotes ℓ_2 normalisation, $\mathcal{K}_+$ contains Known categories with at least one such occurrence, and o indexes warm-up occurrences. The unit-threshold smooth-L1 penalty is $\psi(u) = u^2/2$ for $|u| < 1$ and $\psi(u) = |u| - 1/2$ otherwise. If $\mathcal{K}_+$ is empty, the anchor loss is zero. Matching these similarities preserves each image's relation to the Known reference directions.

Neighbour relations extend this reference to similarities among images, including unlabelled ones. Let $\mathcal{E}$ contain the directed pairs from each instance to its ten nearest valid batch neighbours under teacher cosine similarity. Self-pairs and repeated image identities are excluded; fewer neighbours are used when necessary. We minimise

$$\mathcal{L}_{\text{neighbor}} = \frac{1}{\max(1, |\mathcal{E}|)} \sum_{(i,j) \in \mathcal{E}} \psi\Big(\langle h_i, h_j \rangle - \langle \widehat{h}_i, \widehat{h}_j \rangle \Big). \tag{10}$$

Teacher similarities define both the neighbour graph and its targets, independently of class-label agreement.

4.5 Aligned patch distillation

Aligned patch distillation retains local visual information while category representations adapt. Let $\xi_{it}, \widehat{\xi}_{it} \in \mathbb{R}^d$ be normalised student view-2 and teacher view-1 patch features at position t. For $P = 196$ positions per image,

$$\mathcal{L}_{\text{patch}} = \frac{1}{BP} \sum_{i=1}^{B} \sum_{t=1}^{P} \Big[1 - \langle \xi_{it}, \mathrm{sg}(\widehat{\xi}_{it}) \rangle \Big]. \tag{11}$$

Shared geometry establishes spatial correspondence, while photometric perturbations impose agreement across appearance changes. Matching all visible patches preserves local representations without object annotations or one-category-per-patch assignments.

4.6 Training objective and implementation

Supervised contrastive loss $\mathcal{L}_{\text{supcon}}$ links labelled examples sharing a valid Known positive. Instance contrastive loss $\mathcal{L}_{\text{InfoNCE}}$ links the two views of each sampled instance. Both operate on normalised projection features, with exact reductions in Appendix A.2. Define $\bar{p}_c = (2B)^{-1} \sum_{a=1}^{2} \sum_{i \in \mathcal{B}_L \cup \mathcal{B}_U} p_{ic}^{(a)}$ and normalised marginal mass $\omega_c = \bar{p}_c / \sum_{c'=1}^{C} \bar{p}_{c'}$. Then

$$\mathcal{L}_{\text{bal}} = \sum_{c=1}^{C} \omega_c \log(C\omega_c), \qquad \mathcal{L}_{\text{base}} = 0.35(\mathcal{L}_K + \mathcal{L}_{\text{supcon}}) + 0.65(\mathcal{L}_U + \mathcal{L}_{\text{InfoNCE}} + 0.1\mathcal{L}_{\text{bal}}). \tag{12}$$

Table 2: **ML-GCD dataset partitions and image composition.** Known-dev images are Known-only and excluded from optimisation. Composition counts use benchmark annotations, including VOC difficult objects, before validity filtering for subgroup evaluation (Appendix C).

(a) Categories and image partitions

	Categories		Images			
Dataset	Known	Novel	Labelled	Unlabelled	Known-dev	Test
VOC2007	10	10	836	4,175	0	4,952
MIRFlickr	12	12	1,571	17,701	392	4,917
COCO2014	40	40	7,956	72,136	1,989	40,137

(b) Image composition

	Unlabelled images			Test images		
Dataset	Known-only	Novel-only	Mixed	Known-only	Novel-only	Mixed
VOC2007	835	2,009	1,331	1,602	2,051	1,299
MIRFlickr	1,963	2,996	12,742	967	772	3,178
COCO2014	9,946	10,739	51,451	9,613	5,526	24,998

Here $\mathcal{L}_{\text{base}}$ collects the category-learning terms. The balance term encourages category usage across the batch while leaving individual predictions independent. We clamp $\bar{p}_c$ below at 10^{-8} and apply no explicit cardinality penalty. The complete objective is

$$\mathcal{L}_{\text{PaD}} = \mathcal{L}_{\text{base}} + \rho_e \left[\mathcal{L}_{\text{K-KD}} + \tfrac{1}{2}(\mathcal{L}_{\text{anchor}} + \mathcal{L}_{\text{neighbor}} + \mathcal{L}_{\text{patch}})\right]. \tag{13}$$

For one-based epoch e, $\rho_e = \min(1, \max(0, (e-30)/10))$. The first 30 epochs prepare the teacher head and Known anchors; fixed-reference transfer starts at epoch 31 and reaches full weight at epoch 40. We train the student category head, contrastive projection head, and final backbone block for 200 epochs, then evaluate the final checkpoint. All fixed-reference targets remain detached, and the teacher is discarded at inference.

5 Experiments

5.1 Experimental setup

Datasets and evaluation. We evaluate PaD on VOC2007, MIRFlickr, and COCO2014 (Everingham et al., 2010; Huiskes & Lew, 2008; Lin et al., 2014), with category counts, image partitions, and category composition in Table 2. VOC uses trainval/test; MIR uses a deterministic 80/20 image split; COCO uses train2014/val2014. MIR and COCO reserve Known-only development images outside optimisation. Appendix C details split construction and annotation masks.

Evaluation follows the global alignment and metric definitions in Section 3, with implementation details in Appendix C.

Training. Image-training uses DINO-v1 ViT-B/16 with a trainable final block, aligned 224-pixel views, and batch size 64. Single-stage models use 200 epochs of FP32 SGD, initial learning rate 0.1, momentum 0.9, weight decay 10^{-4}, and cosine decay. We report mean and sample SD over final checkpoints for seeds 0, 1, and 2. COCO follows the checkpoint-recovery protocol in Appendix B.1; Appendix B specifies the study design.

Comparisons. Our ML-GCD adaptations include ML-SimGCD with Bernoulli classification and cross-view distillation (Wen et al., 2023), and MulSupCon with momentum-queue, per-positive-class pretraining followed by classifier transfer and the discovery adapter (Zhang & Wu, 2024). MulSupCon uses 150 pretraining and 50 classification epochs, with different optimisation and unlabelled exposure (Appendix B). All scores come from these implementations.

The *Category-learning control (ASL)* retains PaD's independent category predictions and sparse cross-view targets, using ASL for valid Known annotations and selected targets (Ridnik et al.,

Table 3: **Multi-label category discovery on three datasets.** All/Known/Novel mAP (%, mean ± sample SD, three seeds). All image-training methods use DINO-v1 ViT-B/16 and 200 total epochs. MulSupCon uses a two-stage schedule; COCO follows the checkpoint-recovery protocol.

Method	All ↑	Known ↑	Novel ↑
VOC2007: 10 Known / 10 Novel			
ML-SimGCD (Bernoulli)	24.34 ± 0.47	41.89 ± 1.00	6.80 ± 0.07
Category-learning control (ASL)	59.28 ± 0.84	64.62 ± 1.91	53.95 ± 2.23
MulSupCon framework + adapter	31.98 ± 1.73	49.23 ± 1.00	14.72 ± 4.46
PaD	63.32 ± 5.35	71.71 ± 10.10	54.93 ± 0.73
MIRFlickr: 12 Known / 12 Novel			
Category-learning control (ASL)	46.07 ± 1.09	54.81 ± 4.27	37.34 ± 2.64
PaD	50.92 ± 0.59	60.20 ± 6.35	41.64 ± 6.58
COCO2014: 40 Known / 40 Novel			
Category-learning control (ASL)	46.89 ± 1.69	55.62 ± 1.44	38.16 ± 2.65
PaD	48.11 ± 0.82	56.17 ± 1.55	40.05 ± 0.73

2021). It omits the fixed-reference losses to isolate their contribution and is abbreviated to *Category-learning control* outside Table 3.

5.2 Recognition and Discovery Across Datasets

Overall performance. PaD improves mean All, Known, and Novel mAP over the category-learning control on all three datasets (Table 3). All mAP increases by 4.04, 4.85, and 1.22 percentage points on VOC2007, MIRFlickr, and COCO2014, respectively; Novel mAP increases by 0.98, 4.30, and 1.89 points. These gains support the effectiveness of fixed visual guidance for multi-label category discovery.

The individual runs reveal how these average gains vary with training. Novel mAP improves in two of three paired seeds on each dataset. MIR and COCO have paired Novel differences of $+5.07/+7.97/-0.12$ and $+1.98/+4.27/-0.56$ points, respectively. VOC Known mAP has a sample SD of 10.10 points.

Set prediction. PaD improves All macro/micro F1 from 38.14/39.82 to 54.01/54.10 on VOC, 38.76/42.85 to 43.96/49.75 on MIR, and 28.44/29.85 to 35.79/39.61 on COCO (Table 7). Cardinality MAE decreases from 2.19 to 0.80 on VOC and 2.36 to 1.92 on MIR. These gains complement improved category ranking; Section 5.4 examines mixed-image recognition.

Adapted training strategies. On VOC2007, PaD achieves 54.93 Novel mAP, compared with 6.80 and 14.72 for ML-SimGCD and MulSupCon. These adaptations compare complete training strategies with different target construction, discovery-loss sample scope, and representation-learning stages.

Frozen-feature methods. A complementary frozen-DINO panel evaluates UPS, CAP, Bernoulli-PIM, MCUL, and DMNL (Rizve et al., 2021; Xie et al., 2023; Chiaroni et al., 2023; Huang et al., 2021; Zhu et al., 2017). Appendix B.2 reports these method-specific adaptations and their aggregate results separately from PaD's representation fine-tuning.

5.3 Contributions of Fixed Visual Guidance

We ablate each distillation module on VOC while retaining the training horizon, split, and paired initial states. Table 4 shows complementary effects on Known recognition and Novel discovery.

Global and local feature constraints. Relation preservation raises mean Novel mAP by 1.34 points, with a positive effect in all three paired seeds. Aligned patch distillation contributes 1.32 points on average, with paired differences of $+0.22$, -0.45, and $+4.19$. Thus both constraints improve average discovery performance, with a more variable local contribution.

Table 4: **PaD module ablations on VOC.** mAP (%, mean ± sample SD), using three paired seeds. Each variant removes one module from the fixed full-model configuration.

Configuration	All ↑	Known ↑	Novel ↑
PaD	63.32 ± 5.35	71.71 ± 10.10	54.93 ± 0.73
Without Known distillation	59.76 ± 3.69	63.01 ± 3.36	56.51 ± 4.57
Without relation preservation	59.52 ± 4.81	65.44 ± 9.36	53.59 ± 0.35
Without patch distillation	61.53 ± 2.64	69.45 ± 7.96	53.61 ± 3.16

Table 5: **Mixed-image recognition.** Percentages (mean ± sample SD, three seeds) on fully annotated mixed images. Joint hit requires at least one correct prediction from each category group.

Dataset: method	Joint hit ↑	Precision ↑	Novel recall ↑
VOC: Category-learning control	27.36 ± 0.91	33.85 ± 1.26	39.90 ± 4.30
VOC: PaD	13.08 ± 1.65	43.19 ± 3.36	37.03 ± 1.08
MIR: Category-learning control	37.10 ± 5.60	40.94 ± 1.93	39.36 ± 3.79
MIR: PaD	48.53 ± 1.09	52.67 ± 3.78	45.52 ± 1.19
COCO: Category-learning control	40.52 ± 1.13	21.34 ± 2.66	48.20 ± 2.64
COCO: PaD	27.53 ± 0.45	38.42 ± 0.20	51.82 ± 0.89

Known-prediction transfer. Known distillation raises mean Known mAP by 8.71 points and Novel macro/micro F1 by 8.91/8.94 points, while reducing Novel mAP by 1.58 points. This contrast clarifies its role: transferring Known predictions changes the balance between recognition, ranking, and thresholded decisions. The full model combines this prediction-level guidance with the global and local representation constraints.

5.4 Prediction under mixed category composition

Mixed images test simultaneous Known and Novel recognition. We select images with fully valid annotations and at least one positive from each group. *Joint hit* measures the fraction with at least one correct Known and one correct Novel prediction. Precision pools correct positive predictions across categories; Novel recall pools recovered Novel positives.

Precision and false positives. PaD increases mixed-image prediction precision on all three datasets (Table 5). On fully valid Known-only images, Novel false positives also decrease by 0.98 and 0.79 per image on VOC and MIR, indicating fewer spurious predictions alongside the aggregate F1 gains.

Joint detection and recall. Recall changes depend on the dataset. On MIR, mean joint hit and Novel recall increase by 11.43 and 6.16 percentage points, respectively, with positive gains in every seed. Joint hit decreases by 14.28 points on VOC. On COCO, Novel recall increases by 3.62 points, but Known recall decreases by 33.14 points and joint hit by 12.99 points, with the latter decline present in every seed. These findings distinguish gains in prediction precision from joint category recovery.

6 Conclusion

We introduced ML-GCD and PaD, coupling category discovery with fixed guidance that transfers Known predictions and preserves global and local visual structure. Experiments across three datasets show gains in mean All and Novel mAP, aggregate F1, and mixed-image precision over the category-learning control. Ablations identify complementary roles for prediction transfer and representation preservation. The setting, algorithm, and evaluation protocol support further research on multi-label category discovery.

AI USE STATEMENT

Generative AI tools assisted with language polishing, grammar checking, interpretation of existing experimental results, and preparation of schematic figures.

REPRODUCIBILITY STATEMENT

The appendices specify data construction, objectives, method adaptations, and evaluation, and report aggregate results for image-training and frozen-feature comparisons. The source package includes the LaTeX compilation dependencies. Numerical records and prediction-audit provenance are described in Appendix C.1; the manuscript source package does not include the training code or prediction arrays. COCO checkpoint recovery is specified in Appendix B.1.

ETHICS STATEMENT

This study uses existing visual datasets under their applicable access and redistribution conditions. The task and method figures contain selected VOC2007 examples.

REFERENCES

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 (CVPR)*, pp. 16880–16889, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cao_Solving_the_Catastrophic_Forgetting_Problem_in_Generalized_Category_Discovery_CVPR_2024_paper.html.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

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*, 2021. URL https://arxiv.org/abs/2104.14294.

Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 522–531, October 2019a. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Chen_Learning_Semantic-Specific_Graph_Representation_for_Multi-Label_Image_Recognition_ICCV_2019_paper.html.

Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5177–5186, June 2019b. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Chen_Multi-Label_Image_Recognition_With_Graph_Convolutional_Networks_CVPR_2019_paper.html.

Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1729–1739, 2023.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. URL https://www.robots.ox.ac.uk/~vgg/publications/2010/Everingham10/everingham10.pdf.

Jun Huang, Linchuan Xu, Jing Wang, Lei Feng, and Kenji Yamanishi. Discovering latent class labels for multi-label learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 3058–3064, 2020. doi: 10.24963/ijcai.2020/423. URL https://www.ijcai.org/proceedings/2020/423.

Jun Huang, Linchuan Xu, Kun Qian, Jing Wang, and Kenji Yamanishi. Multi-label learning with missing and completely unobserved labels. *Data Mining and Knowledge Discovery*, 35:1061–1086, 2021. doi: 10.1007/s10618-021-00743-x. URL https://link.springer.com/article/10.1007/s10618-021-00743-x.

Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, 2008. URL https://press.liacs.nl/mirflickr/mirflickr.pdf.

Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16478–16488, June 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Lanchantin_General_Multi-Label_Image_Classification_With_Transformers_CVPR_2021_paper.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2014/09/LinECCV14coco.pdf.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1904.05068.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ridnik_Asymmetric_Loss_for_Multi-Label_Classification_ICCV_2021_paper.html.

Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. ML-Decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 32–41, January 2023. URL https://openaccess.thecvf.com/content/WACV2023/html/Ridnik_ML-Decoder_Scalable_and_Versatile_Classification_Head_WACV_2023_paper.html.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2101.06329.

Luyao Tang, Kunze Huang, Chaoqi Chen, Yuxuan Yuan, Chenxin Li, Xiaotong Tu, Xinghao Ding, and Yue Huang. Dissecting generalized category discovery: Multiplex consensus under self-deconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 297–307, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Tang_Dissecting_Generalized_Category_Discovery_Multiplex_Consensus_under_Self-Deconstruction_ICCV_2025_paper.html.

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*, 2022. URL https://arxiv.org/abs/2201.02609.

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, June 2025. URL https://openaccess.thecvf.com/content/

CVPR2025/html/Wang_GET_Unlocking_the_Multi-modal_Potential_of_CLIP_for_Generalized_Category_CVPR_2025_paper.html.

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

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*, 2023. URL https://arxiv.org/abs/2211.11727.

Ming-Kun Xie, Jia-Hao Xiao, Hao-Zhe Liu, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Class-distribution-aware pseudo labeling for semi-supervised multi-label learning. *arXiv preprint arXiv:2305.02795*, 2023. URL https://arxiv.org/abs/2305.02795.

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 (ICCV)*, pp. 405–414, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xu_A_Hidden_Stumbling_Block_in_Generalized_Category_Discovery_Distracted_Attention_ICCV_2025_paper.html.

Pingyue Zhang and Mengyue Wu. Multi-label supervised contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16786–16793, 2024. doi: 10.1609/aaai.v38i15.29619. URL https://ojs.aaai.org/index.php/AAAI/article/view/29619.

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, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_PromptCAL_Contrastive_Affinity_Learning_via_Auxiliary_Prompts_for_Generalized_Novel_CVPR_2023_paper.html.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.07832.

Ke Zhu and Jianxin Wu. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 184–193, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zhu_Residual_Attention_A_Simple_but_Effective_Method_for_Multi-Label_Recognition_ICCV_2021_paper.html.

Xuelin Zhu, Jian Liu, Jiuxin Cao, and Bing Wang. MambaML: Exploring state space models for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4743–4753, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_MambaML_Exploring_State_Space_Models_for_Multi-Label_Image_Classification_ICCV_2025_paper.html.

Yue Zhu, Kai Ming Ting, and Zhi-Hua Zhou. Discover multiple novel labels in multi-instance multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. doi: 10.1609/aaai.v31i1.10838. URL https://ojs.aaai.org/index.php/AAAI/article/view/10838.

A NOTATION AND AUXILIARY LOSS DEFINITIONS

Unless otherwise specified, i, j index instances; c indexes all outputs; r indexes anonymous Novel outputs; a, b index views; t indexes patch positions; and e indexes epochs. In Equations 2 and 17, j indexes true Novel categories and ranking positions, respectively. The indexed probability $p_{ic}^{(a)}$ abbreviates $p_c(x_i^{(a)})$. A view-free p_{ic} denotes the inference score before evaluation alignment. After alignment, p_{ic}^{eval} is expressed in true-category order.

Table 6: **Notation used in the manuscript.** Sets, masks, probabilities and feature representations are kept distinct.

Symbol	Meaning
$\mathcal{K}, \mathcal{N}, \mathcal{C}$; K, R, C	Known, Novel and complete category sets; their cardinalities, with $C = K + R$.
$\mathcal{D}_L, \mathcal{D}_U, \mathcal{D}_{\text{dev}}$	Labelled training, unlabelled training and optional Known validation sets.
n_L, n_U, n_E	Numbers of labelled, unlabelled and evaluation instances.
x_i, y_{ic}, v_{ic}	Instance, binary category-presence target and annotation-validity indicator.
S	Labelled-set membership indicator used in the selection rule.
$\mathcal{B}_L, \mathcal{B}_U$; B_L, B_U, B	Labelled/unlabelled batch index sets and counts; $B = B_L + B_U$.
a, b; $x_i^{(a)}$	View indices in $\{1, 2\}$; augmented input view.
θ, ϕ	Trainable student parameters and frozen visual-teacher parameters.
f_θ, f_ϕ; d, d_g	Student/teacher visual encoders; backbone dimension 768 and contrastive projection dimension 256.
$h_i^{(a)}, h_i, \widehat{h}_i$	Normalised student global feature, its view-2 abbreviation, and weak-view fixed-teacher global feature.
$g_i^{(a)}$	Normalised contrastive projection feature; distinct from the classifier feature $h_i^{(a)}$.
w_c, α_c, s_c, b_c	Unit classifier direction, raw scale parameter, positive scale and bias.
$z_{ic}^{(a)}, p_{ic}^{(a)}, \widehat{p}_{ic}$	Student logit, student presence probability, and fixed Known-teacher probability.
m, γ_+, γ_-	ASL negative margin and positive/negative focusing exponents.
$Z_U^{(a)}, Q^{(a)}, \boldsymbol{\mu}$	Unlabelled logit matrix, balanced assignment-score matrix, and running logit centre.
τ_{assign}, q	Assignment temperature and rank cutoff for ignored pseudo-target entries.
$\widetilde{y}_{ic}^{(a)}, M_{ic}^{(a)}$	Detached binary pseudo-target and its selection mask; M is not the annotation mask v.
$\mathcal{W}_c, \mathcal{K}_+, \kappa_c$	Warm-up positive-occurrence multiset, active Known anchor classes and normalised class anchor.
$\mathcal{E}, \psi$	Directed teacher-neighbour edge set and unit-threshold smooth-L1 penalty.
$\xi_{it}, \widehat{\xi}_{it}, P$	Student/teacher aligned patch features and number of patch positions ($P = 196$).
$\bar{p}_c, \omega_c$	Batch-mean presence probability and normalised category marginal.
ρ_e	Epoch-dependent weight of all fixed-reference transfer losses.
$\mathcal{L}_K, \mathcal{L}_U$	Known annotation loss and sparse cross-view category loss.
$\mathcal{L}_{\text{supcon}}, \mathcal{L}_{\text{InfoNCE}}, \mathcal{L}_{\text{bal}}$	Label-supervised contrastive, instance contrastive and marginal-balance losses.
$\mathcal{L}_{\text{K-KD}}, \mathcal{L}_{\text{anchor}}$	Fixed Known-prediction and Known-anchor losses.
$\mathcal{L}_{\text{neighbor}}, \mathcal{L}_{\text{patch}}$	Neighbour-relation and aligned patch-transfer losses.
$\mathcal{L}_{\text{base}}, \mathcal{L}_{\text{PaD}}$	Category-learning objective and complete PaD objective.
$A_{rj}, \mathfrak{S}_R, \pi^*$	Pairwise Novel AP, permutations of Novel categories and selected global matching.

Symbol	Meaning
$\mathcal{I}, \mathcal{I}_L, \mathcal{P}_\ell$	All batch-view pairs, labelled batch-view pairs, and contrastive positives for pair ℓ.
$\tau_{\text{sup}}, \tau_{\text{inst}}$	Supervised and instance contrastive temperatures (0.07 and 0.1).
$\widetilde{p}_{ic}^{(a)}, \tau_{\text{ctrl}}(e)$	Targets and scheduled teacher temperature of the historical Bernoulli control.
σ_{blur}	Gaussian-blur standard deviation; distinct from the sigmoid function σ.
$n_c^+, n_c^v, y_{(k)c}$	Valid positive count, valid instance count and label at score rank k for evaluation category c.
$\mathcal{P}_r^+, \mathcal{P}_r^-, \mathcal{L}_{\text{fixed}}$	Positive/negative pseudo-target index sets and fixed-branch loss of the regional variant.

A.1 Implementation details

The category classifier uses $d = 768$ backbone features. A three-layer projection head maps unnormalised backbone features to $d_g = 256$ dimensions for contrastive learning. Classifier scales initialise at 10 and biases at $\log(0.075/0.925)$, independently of hidden labels. We use $\sigma(z) = (1 + e^{-z})^{-1}$ and $\text{softplus}(z) = \log(1 + e^z)$.

The VOC rank cutoff is $q = 5$. The retained source gives two cross-dataset specifications: the main method uses a fixed cutoff of five, whereas preparatory configurations use $q = \max(5, \lceil C/4 \rceil)$. The completed result tables do not identify which cutoff was used for each MIR/COCO run. This configuration detail remains unverified in the retained manuscript materials.

For ASL, focusing weights are detached when differentiating.

A.2 Contrastive objectives

The projection head maps unnormalised backbone global features to d_g-dimensional vectors; their normalised versions are $g_i^{(a)}$. Let $\mathcal{I} = \{(i, a) : i \in \mathcal{B}_L \cup \mathcal{B}_U,\ a \in \{1, 2\}\}$ and let $\mathcal{I}_L$ restrict this set to labelled indices. For a set of batch-view pairs $\mathcal{T}$, define a nonempty positive set $\mathcal{P}_\ell \subseteq \mathcal{T} \setminus \{\ell\}$ for each $\ell \in \mathcal{T}$. With temperature $\tau > 0$, the contrastive reduction is

$$\mathcal{L}_{\text{ctr}}(\mathcal{T}, \{\mathcal{P}_\ell\}, \tau) = -\frac{1}{|\mathcal{T}|} \sum_{\ell \in \mathcal{T}} \frac{1}{|\mathcal{P}_\ell|} \sum_{k \in \mathcal{P}_\ell} \log \frac{\exp(\langle g_\ell, g_k \rangle / \tau)}{\sum_{u \in \mathcal{T} \setminus \{\ell\}} \exp(\langle g_\ell, g_u \rangle / \tau)}. \tag{14}$$

Here ℓ, k, u index batch-view pairs and $g_{(i,a)} = g_i^{(a)}$. For $\mathcal{L}_{\text{InfoNCE}}$, use $\mathcal{T} = \mathcal{I}$, $\tau = \tau_{\text{inst}} = 0.1$, and the other view of the same sampled instance as the sole positive. For $\mathcal{L}_{\text{supcon}}$, use $\mathcal{T} = \mathcal{I}_L$ and $\tau = \tau_{\text{sup}} = 0.07$. A pair is positive when its instances share at least one category with $v_{ic} y_{ic} = v_{jc} y_{jc} = 1$ for $c \le K$; the other view of the same instance is also positive. Exact self-pairs are excluded from numerator and denominator. Empty labelled batches contribute zero supervised contrastive loss.

A.3 Assignment normalisation and centre update

A common maximum is subtracted before elementwise exponentiation in Equation 6. In Equation 6, start with the transposed nonnegative matrix $W \in \mathbb{R}^{C \times B_U}$ and divide it by its total mass. In each of three iterations, divide every category row by its row sum and by C, then divide every instance column by its column sum and by B_U. Multiply by B_U and transpose to obtain $Q^{(a)}$. Denominators are clamped below at 10^{-12}. The output rows have unit mass, with approximate category balance after the finite normalisation iterations. After the student update, the centre is updated as

$$\boldsymbol{\mu} \leftarrow 0.9\boldsymbol{\mu} + 0.1 \left[\frac{1}{2B} \sum_{a=1}^{2} \sum_{i \in \mathcal{B}_L \cup \mathcal{B}_U} \text{sg}(z_i^{(a)}) \right]. \tag{15}$$

The vector $z_i^{(a)} \in \mathbb{R}^C$ contains all category logits for that view. The centre starts at zero.

B CURRENT COMPARISON PROTOCOLS AND ADDITIONAL RESULTS

The VOC/MIR image-training panels contain 29 runs: 12 VOC method comparisons, nine VOC module ablations, six MIR control/PaD runs, and two single-seed internal interventions. Each run reaches its specified final checkpoint. The COCO and frozen-feature studies are reported separately below. Earlier single-seed studies appear in the later appendices.

Study design. Each dataset uses one fixed Known/Novel class partition and a supplied Novel category count. Reported seed variation reflects training variation within that partition. Earlier VOC test observations informed development; the VOC results are therefore exploratory. Novel identities are aligned post hoc for evaluation using a single global mapping. The method adaptations retain their stated optimisation schedules; COCO runs use the checkpoint-recovery protocol below.

Adaptation scope. ML-SimGCD uses BCE Known supervision and Bernoulli cross-view distillation over the mixed training batch. The category-learning control uses ASL for valid Known annotations and balanced sparse targets on unlabelled instances. Its comparison with ML-SimGCD changes both the loss and its sample scope. PaD jointly optimises these category-learning terms and the fixed-reference constraints defined in the method section. Dual and loss-only MulSupCon are single-seed internal interventions. Dual extends Novel distillation, while loss-only MulSupCon changes the supervised contrastive term within the shared single-stage route. The full momentum-queue adaptation is specified separately below.

Control naming. The current category-learning control appears as ASL, ASL adapter, ASL control, or ASL + discovery adapter in earlier run records and summaries. These names refer to the same sparse-discovery configuration. The historical *ASL-supervised* rows in the later single-seed panels use Bernoulli cross-view targets and denote a different configuration; their labels are retained.

PaD module ablations. The relation ablation in Table 4 removes anchors and neighbours together, evaluating their combined effect. All PaD module ablations retain the training horizon, class partition, and paired initial states of the full-model comparison.

MulSupCon framework adaptation. The full adaptation uses the source implementation at commit `42b23028785203123b673b8a28f5391b9d9726be`. It retains momentum updates (0.999), a 2,048-entry queue, 256-dimensional projection, temperature 0.1, per-positive-class supervision, and a linear classifier initialised from the trained query backbone. Queue-occupancy masking and partial-label numerical stabilisation are explicit extensions. The 200-epoch budget comprises 150 labelled-only pretraining epochs and 50 classification epochs with balanced labelled/unlabelled sampling. These stages perform 11,700 and 3,900 updates, respectively. Classification uses 0.35 times Known BCE plus 0.65 times the sum of the shared sparse unlabelled loss and 0.1 times the distribution-balance loss. It includes no PaD distillation or additional classifier-stage supervised contrastive loss.

Both stages use Adam with learning rate 4×10^{-4} and weight decay 10^{-5}. The backbone learning-rate multiplier is 1.0 during pretraining and 0.1 during classification; no scheduler is used. Pretraining draws 748,800 images with replacement from the labelled subset; classification draws 249,600 images, approximately half unlabelled. Consequently, unlabelled exposure and classifier initialisation differ from the single-stage controls. These choices define the ML-GCD adaptation, with the recorded schedule used for every reported seed.

Supervision probe protocol and results. The supervision probe uses frozen DINO features from 836 labelled VOC images, with 515 fitting, 166 selection, and 155 audit images. Five episodes each treat two of the ten Known categories as simulated Novel categories. Each condition performs 200 head-training epochs (1,800 updates) per episode, with identical paired inputs and initial states; only the fitting labels for the simulated Novel categories differ. The seed-0 pair uses the same GPU and initial logits; hardware equivalence across seeds remains unverified. We first average the scores across the five episodes within each seed, then report the mean and sample SD across the three seeds. Across three paired training seeds, revealing their fitting labels raises simulated-Novel mAP from 10.36 ± 0.63 to 39.45 ± 1.23. This probe measures the value of category supervision in the fixed-feature classifier and complements PaD's use of the available Known labels.

Table 7: **Set prediction in the current panels.** F1 is in percent; MAE is labels per image. Values are mean ± sample SD across three seeds except Dual and the loss-only MulSupCon intervention (one seed each). The MulSupCon framework and loss-only intervention are reported separately.

Dataset: configuration	Macro F1 ↑	Micro F1 ↑	MAE ↓
VOC: Category-learning control	38.14 ± 1.03	39.82 ± 1.29	2.19 ± 0.08
VOC: Dual	51.36	48.53	0.90
VOC: ML-SimGCD	4.84 ± 0.83	3.96 ± 0.63	1.387 ± 0.005
VOC: MulSupCon loss only	36.18	37.52	2.21
VOC: PaD	54.01 ± 5.58	54.10 ± 4.34	0.80 ± 0.29
VOC: PaD w/o Known KD	38.94 ± 1.02	40.44 ± 1.27	2.25 ± 0.06
VOC: PaD w/o relations	50.55 ± 5.32	51.04 ± 4.36	0.90 ± 0.21
VOC: PaD w/o patch distillation	52.30 ± 3.52	52.33 ± 1.16	0.88 ± 0.13
MIR: Category-learning control	38.76 ± 0.18	42.85 ± 0.65	2.36 ± 0.12
MIR: PaD	43.96 ± 1.42	49.75 ± 2.55	1.92 ± 0.06
VOC: MulSupCon framework	18.94 ± 4.99	20.66 ± 8.71	2.94 ± 0.98

Table 8: **Frozen-feature adaptations on VOC and COCO.** mAP (%, mean ± sample SD, three seeds). Parentheses denote SD below 0.01. All methods use frozen DINO features; their optimisation differs from the image-training panel.

Adaptation	All	Known	Novel
VOC: Known-BCE + G0-Novel	22.72 ± 0.11	38.12 ± 0.13	7.32 ± 0.20
VOC: UPS-Known + G0-Novel	20.36 ± 0.09	31.84 ± 0.15	8.87 ± 0.10
VOC: CAP-Known + G0-Novel	38.45 ± 0.80	69.58 ± 1.49	7.32 ± 0.20
VOC: Bernoulli-PIM	21.49 ± 1.95	22.99 ± 1.20	20.00 ± 3.10
VOC: MCUL-mask-Nystrom	58.92 ± 1.88	78.17 ± 1.20	39.68 ± 2.78
VOC: DMNL-regions-mask	28.77 ± 2.11	46.65 ± 2.86	10.90 ± 1.40
COCO: Known-BCE + G0-Novel	21.37 ± 0.01	40.48 ± 0.03	2.26 (<0.01)
COCO: UPS-Known + G0-Novel	18.05 ± 0.01	33.78 ± 0.02	2.32 (<0.01)
COCO: CAP-Known + G0-Novel	26.27 ± 0.74	50.27 ± 1.48	2.26 (<0.01)
COCO: Bernoulli-PIM	13.74 ± 0.10	18.24 (<0.01)	9.24 ± 0.19
COCO: MCUL-mask-Nystrom	24.38 ± 0.92	34.08 ± 1.68	14.68 ± 0.77
COCO: DMNL-regions-mask	12.19 ± 1.49	20.64 ± 2.91	3.74 ± 0.09

Active-update timing. The timing harness uses batch size 64, FP32, four data workers, and six CPU threads on a Tesla V100-PCIE-32GB. Both arms load mature checkpoints, run ten warm-up and 50 measured updates, and retain the student, EMA teacher, fixed backbone, and Known head. Active-update timing covers host-to-device transfer, objective evaluation, backward/update operations, and periodic logging under shared teacher infrastructure. Teacher preparation, loader waiting, input-audit hashing, and the cost difference relative to a teacher-free implementation are outside the measured scope. The category-learning control and PaD take 0.749 and 0.741 seconds per measured update, with peak allocated memory of 3.22 GiB for both. These latencies are descriptive measurements from one timing allocation per arm.

B.1 COCO CHECKPOINT RECOVERY

All six planned runs were interrupted by infrastructure wall-clock limits. The continued runs restore model, optimiser momentum, scheduler, centre, distillation state, and teacher optimiser. The original checkpoints lack random-number-generator and sampler states, which are reinitialised under the recovery protocol. All runs complete the same total training horizon under this continuation protocol. Reinitialising RNG and sampler states changes the trajectory relative to uninterrupted training. COCO uses eight workers, compared with four for VOC/MIR.

B.2 FROZEN-FEATURE TRANSFERS OF RELATED METHODS

Shared feature and annotation boundary. The DINO-v1 ViT-B/16 encoder remains frozen. Global-feature methods use deterministic 224-pixel views and training-centred, normalised CLS

features. Training includes 5,011 VOC or 80,092 COCO instances. Only permitted Known annotations from $\mathcal{D}_L$ enter optimisation or model selection. Labels in $\mathcal{D}_U$ and all Novel annotations remain hidden. The source objectives are adapted to this mask; absent annotations are not supplied as negative targets. The fixed features, head budgets, and feature-space perturbations differ from the image-training panels.

Known-BCE, UPS, and CAP with shared Novel discovery. The shared head maps 768 features to category logits, with a bias, logit scale 25, and feature dropout 0.3. Training uses 200 head epochs, batches with 64 labelled and up to 192 unlabelled instances, and SGD with learning rate 0.03, momentum 0.9, weight decay 5×10^{-4}, and cosine decay. G0-Novel applies Bernoulli cross-view learning to unlabelled Novel outputs. Its student temperature is 0.2; the teacher temperature decreases from 0.4 to 0.25 over 30 epochs. Centre momentum is 0.9, target cardinality is 1.5, and distribution balancing has weight 0.1. Two feature-dropout views replace image augmentations. Known-BCE adds no Known pseudo-labels.

UPS retains uncertainty-filtered positive and negative pseudo-label selection (Rizve et al., 2021). Ten Monte Carlo dropout passes estimate uncertainty. Positive/negative confidence thresholds are 0.5/0.05 and uncertainty thresholds are 0.05/0.005; negative prediction uses temperature 2. Five rounds of 40 epochs reset the head, optimiser, and centre. The first round uses no pseudo-labels, and initial selection uses a per-class floor of 25. CAP retains class-frequency-based thresholds, an EMA predictor with decay 0.9997, 12 warm-up epochs, and per-unlabelled-epoch refresh (Xie et al., 2023). Its ASL parameters are 4/0 with clipping 0.05; positive/negative selection proportions are 1/0.998. Frequencies are estimated from permitted Known labels only. The composition restriction introduces a potential bias in these estimates. Novel class frequencies are not inferred from hidden targets. SGD with cosine decay replaces the source optimiser schedule.

Bernoulli-PIM. The PIM adaptation retains information maximisation and label-limited weight selection, while replacing categorical entropy with independent Bernoulli entropies (Chiaroni et al., 2023). The objective combines negative marginal entropy, weighted conditional entropy on $\mathcal{D}_U$, and masked Known BCE on $\mathcal{D}_L$. A linear head with bias and scale 25 uses Adam at 0.001 with weight decay 0.01 for 1,000 full-batch updates. Unsupervised candidates cover weights 0.05 to 1 in increments of 0.05; selection among 0.1 to 1 uses Hungarian-aligned macro AP on training Known annotations only. Initialisation uses 100 iterations of training-feature k-means++, Known feature means, and training-median intercepts. This formulation transfers PIM's information-maximisation principle to independent multi-label outputs under ML-GCD supervision.

MCUL and DMNL. MCUL-mask-Nystrom implements the source latent-label objective with annotation masking, Gaussian feature/label similarities, and a rank-256 Nyström approximation (Huang et al., 2021). Training-median bandwidths and uniform training landmarks require no hidden labels. The five objective coefficients are 1, 100, 10, 5, and 0.01. Optimisation uses 2,000 alternating updates with projected latent labels, proximal predictor updates, and a pseudoinverse label-relation update. The low-rank approximation, solver, and graph-dimension correction are explicit modifications to the source formulation.

DMNL-regions-mask retains bag prediction, instance clustering, dual updates, orthogonality, and penalty ramping (Zhu et al., 2017). Sixteen regions are pooled on a 4×4 grid from the final 14×14 DINO patch map; no object boxes or masks are used. Regional features use normalisation, 200 random Fourier features, and training-based min–max scaling. Each of 500 outer iterations uses five full-batch L-BFGS predictor updates and ten latent-label updates, replacing the source's 50 stochastic predictor and 50 latent updates. The implementation uses the exact softmax Jacobian instead of a diagonal approximation. Weighted regional softmax outputs form bag scores. The reported DMNL results correspond to this regional-feature representation and revised solver schedule.

Comparison scope. The adaptations share the ML-GCD data boundary but retain different predictors, solvers, and score scales. The shared Novel adapter supplies a discovery objective for UPS and CAP. MCUL and DMNL retain their latent-label structure with the stated feature and solver changes. Their fixed optimisation horizons define the reported experiments; source-objective convergence is not established. These comparisons characterise the transferred implementations, separately from PaD's representation fine-tuning.

C PROTOCOL AND REPRODUCIBILITY DETAILS

Dataset construction. The benchmark administrator reads full annotations to create a class split, identify eligible Known-only images, and verify coverage. This administrative use is distinct from training access. The learner receives only valid Known targets for labelled images and no targets for unlabelled images. The category boundary is defined over the benchmark annotation vocabulary, leaving objects outside that vocabulary unspecified. The historical VOC loader retains class names and evaluation metadata in the benchmark manifest, reads annotations, then zeros all unlabelled targets and all Novel target columns before training collation. Annotation isolation in this loader is enforced at the optimisation-tensor level; the loading process retains access to the raw annotations. The model and losses do not use Novel names. The newer MIR/COCO loaders use separate training and evaluation manifests with a stricter field schema; their data-access isolation mechanisms therefore differ.

For VOC, Known classes are aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, and cow. Novel classes are diningtable, dog, horse, motorbike, person, pottedplant, sheep, sofa, train, and tvmonitor. These names document the offline benchmark partition; training uses anonymous Novel output indices. The seed-0 selection labels approximately half of the eligible Known-only images. There are 607 difficult entries in trainval and 619 in test. The evaluation mask removes these entries from both numerator and denominator contributions. Masking operates per label entry, retaining valid labels from the same image.

Table 2 reports the dataset inventory for the current experiments. The Known-only development subset $\mathcal{D}_{\text{dev}}$ is excluded from optimisation; its images are not included in $\mathcal{D}_L$ or $\mathcal{D}_U$.

The composition counts in Table 2 describe the partitions before evaluation masking. VOC counts use XML object presence, including difficult objects, to preserve the Known-only labelling boundary. The fully valid VOC test subset used for composition-specific set metrics contains 4,401 images: 1,579 Known-only, 1,902 Novel-only, and 920 mixed. MIR and COCO have no ignored annotation entries under their presence-label protocols, so their corresponding evaluation groups coincide with the test compositions in the table.

MIRFlickr uses 24 official potential-label annotations, including food; relevant-label files are excluded (Huiskes & Lew, 2008). Hierarchical label groups remain on the same side of the Known/Novel partition. Images are divided by a deterministic custom 80/20 split. COCO uses official train2014 and val2014 boundaries, 80 object-presence labels, and includes crowd instances; repeated instances collapse to one positive label (Lin et al., 2014). Zero-positive images are removed in both cross-dataset builders. A class partition is the first training-coverage-valid candidate among 32 predetermined seed-0 candidates. Test labels are not used to choose a partition. Half of eligible Known-only training images are selected, and 20% of that selection is withheld as Known-dev. These choices differ from the historical VOC protocol and from published multi-label benchmarks.

Composition-dependent selection. If S denotes labelled-set membership, the selection rule is

$$\Pr(S = 1 \mid \|y^{\mathcal{N}}\|_0 > 0) = 0. \tag{16}$$

Optimisation. The backbone is DINO-v1 ViT-B/16 with a three-layer projection head. Trainable backbone parameters belong only to block 11. The head and the last block are trained jointly; earlier parameters remain frozen. The single-stage image-training methods share the initial head scale/prior, data ordering policy, batch size, precision, learning rate, and 200-epoch horizon. VOC runs perform 78 updates per epoch. MIR runs perform 301, yielding 60,200 updates. COCO uses 1,251 updates per epoch and 250,200 in total. The inherited VOC training-pool evaluation is diagnostic; checkpoint evaluation follows the fixed final-epoch rule. Cross-dataset runs evaluate Known-dev every 20 epochs and open the final-test manifest only after optimisation.

Bernoulli control. This historical control uses detached targets $\widetilde{p}_{ic}^{(a)} = \sigma((\text{sg}(z_{ic}^{(a)}) - \mu_c)/\tau_{\text{ctrl}}(e))$, where μ_c is the category component of the running centre and $\tau_{\text{ctrl}}(e)$ is the control teacher temperature at epoch e. Targets are rescaled by $1.5/\sum_{c'=1}^{C} \widetilde{p}_{ic'}^{(a)}$ and clipped to $[10^{-4}, 1 - 10^{-4}]$. Opposite-view student logits are divided by the fixed temperature 0.2 and optimised with binary cross-entropy. The control teacher temperature changes linearly from 0.4 to 0.25 during the first 30

epochs. This rescaling changes probability semantics. The term applies to both labelled and unlabelled images. ML-SimGCD uses BCE for Known labels; ASL-supervised changes that Known loss to ASL. PaD applies rank-selected sparse targets on $\mathcal{D}_U$, whereas these historical controls apply rescaled Bernoulli targets to both labelled and unlabelled images.

Evaluation implementation. First express predictions in true-category order: $p_{ic}^{\text{eval}} = p_{ic}$ for Known categories and $p_{i,K+j}^{\text{eval}} = p_{i,K+(\pi^*)^{-1}(j)}$ for Novel category j. For category c, let n_c^v be the number of valid evaluation instances and $n_c^+ > 0$ their positive count. Sort these instances by descending p_{ic}^{eval}, and let $y_{(k)c}$ be the label at rank k. Ranking AP is

$$\text{AP}_c = \frac{1}{n_c^+} \sum_{k=1}^{n_c^v} y_{(k)c} \left(\frac{1}{k} \sum_{j=1}^{k} y_{(j)c} \right). \tag{17}$$

The inner index j is a ranking position local to this equation. Sorting is stable and tied scores are not grouped. Macro F1 averages categorywise F1 values, whereas micro F1 pools valid true-positive, false-positive and false-negative counts. The valid-entry cardinality error is

$$\text{MAE}_{\text{card}} = \frac{1}{n_E} \sum_{i=1}^{n_E} \left| \sum_{c=1}^{C} v_{ic} \mathbf{1}[p_{ic}^{\text{eval}} \geq 0.5] - \sum_{c=1}^{C} v_{ic} y_{ic} \right|, \tag{18}$$

where $\mathbf{1}[\cdot]$ is the indicator function and n_E is the evaluation-set size. All metrics reuse the same AP-maximising global Novel permutation. Matching on the evaluation set measures recovery up to permutation and introduces a risk of optimistic estimates relative to fixing the mapping on independent data. The AP implementation above resolves score ties by stable image order; tied scores are not grouped. The reported results retain this evaluation convention.

C.1 Numerical provenance

The original experiment records describe prediction-based recomputation of final metrics, including 34 locally retained adaptation arrays and all six COCO prediction/diagnostic files; two adaptations use recorded remote recomputation. The earlier four frozen-feature arrays were recomputed to an absolute discrepancy below 10^{-7}, and recorded audits cover the six distillation and four evidence-assignment variants. All numerical tables in this manuscript are taken from these retained experiment records.

D Complete exploratory numerical results

The historical numerical snapshot, separate from the current three-seed panels, contains 37 completed performance rows: twelve component arms (including the reference), two other primary VOC controls, five earlier VOC interventions, two MIR controls, two balance-gradient arms, six distillation arms, four evidence-assignment arms, and four frozen-feature external adaptations. The first two tables retain the nineteen reference/component/historical VOC rows; later sections give the other groups separately. The fusion reference is counted once even though its disabled-component rerun agrees with the earlier recorded fusion metrics. All rows use one training seed. Source data retain the original fractions; displayed percentages are rounded to two decimals. Rounding maps sufficiently small positive F1 values to 0.00.

Historical interventions. Relaxing the log-domain transport frequency constraint changes All/Known/Novel mAP by $-0.81/-9.00/+7.38$ points relative to its hard-frequency control. Thus its Novel increase is accompanied by a Known decline. The log-domain control and exponent-domain reference are numerically distinct; runtime and floating-point trajectories remain potential contributors to their differences. The overlap-weighted supervised contrastive adaptation has lower All/Known/Novel mAP than the fusion reference. Cross-view negative veto has lower All and Novel mAP than its random-mask control. The contrasts evaluate these component implementations under the stated training conditions.

Table 9: Reference, component, and historical VOC mAP and cardinality results. The rows document different exploratory interventions; causal comparisons require the matched pairs described in the text.

Method	All	Known	Novel	Card. MAE
Fusion reference	54.08	57.58	50.58	2.28
A: variable / relaxed	19.21	24.35	14.06	13.77
B: local query	54.17	68.63	39.71	2.56
B: sparse dictionary	56.53	73.42	39.64	2.33
C: random correspondence	16.08	16.84	15.31	9.84
C: visible correspondence	16.61	17.89	15.33	9.78
D: Known composition	54.10	59.82	48.38	2.16
D: + Novel preservation	56.15	69.95	42.36	2.42
ML-SimGCD adaptation	23.42	39.85	6.99	1.40
ASL-supervised	34.55	60.54	8.57	1.42
Hard transport (log)	57.80	73.05	42.54	4.69
Relaxed transport	56.98	64.05	49.91	1.88
Label-overlap weighting	47.57	52.70	42.44	2.09
Random negative mask	51.72	51.82	51.61	2.11
Cross-view negative veto	48.88	48.27	49.48	2.20
A: fixed / relaxed	34.28	44.92	23.64	8.01
A: fixed / uniform	29.76	40.63	18.90	9.02
A: variable / uniform	17.30	20.81	13.78	11.68
B: dictionary, no penalties	55.05	70.93	39.17	2.43

Table 10: Reference, component, and historical VOC F1 vector, in percent. A/K/N denote All/Known/Novel. Every method uses threshold 0.5 and the same global mapping for all F1 components.

Method	A-macro	A-micro	K-macro	K-micro	N-macro	N-micro
Fusion reference	36.86	37.73	40.43	40.91	33.29	34.56
A: variable / relaxed	14.90	16.09	12.53	12.77	17.27	19.38
B: local query	34.47	36.29	40.26	40.90	28.69	32.14
B: sparse dictionary	36.06	39.17	41.98	42.41	30.15	36.30
C: random correspondence	17.78	19.57	16.42	16.86	19.13	22.31
C: visible correspondence	17.82	19.27	15.91	16.32	19.74	22.38
D: Known composition	36.07	37.65	40.43	41.91	31.71	33.33
D: + Novel preservation	33.91	35.97	38.06	38.79	29.76	32.74
ML-SimGCD adaptation	3.29	2.81	6.58	6.73	0.00	0.00
ASL-supervised	0.04	0.03	0.07	0.07	0.00	0.00
Hard transport (log)	25.92	26.28	31.33	31.68	20.51	21.99
Relaxed transport	43.11	43.11	51.44	51.01	34.77	37.21
Label-overlap weighting	35.19	36.33	42.13	43.34	28.25	29.70
Random negative mask	36.18	37.30	41.03	42.92	31.33	32.05
Cross-view negative veto	36.03	37.12	40.42	41.61	31.64	32.71
A: fixed / relaxed	20.92	21.21	21.76	20.72	20.07	21.65
A: fixed / uniform	18.46	19.66	17.85	18.29	19.07	20.96
A: variable / uniform	15.41	16.71	14.29	14.61	16.53	18.80
B: dictionary, no penalties	35.38	37.50	41.86	42.44	28.89	33.33

E COMPONENT DESIGNS AND EXPERIMENTAL CONTRASTS

Each component is activated independently against the fixed reference; these experiments are separate from PaD.

A. The assignment module optimises a Bernoulli-KL projection with soft penalties for row/cardinality and column/frequency interval violations. The design crosses a fixed row tether near 1.5 with a prediction-derived variable interval, and a uniform frequency tether with a wider prediction-derived frequency interval. All factorial cells share the solver, masking, and numerical iteration budget. These controls use soft interval penalties in place of the reference's hard top-one/top-two selection. The marginal-balance regulariser remains active when the assignment-frequency prior is relaxed. Prediction-derived intervals specify heuristic optimisation bounds; statistical coverage is

unspecified. The box-constrained solver requires a projected-residual convergence criterion; numerical fixtures cover selected diagnostic cases.

B. The patch dictionary uses nonnegative unrolled sparse coding and max-presence pooling. Known atoms are anchored by ordinary Known supervision; an unassigned residual is retained. Reconstruction targets are detached normalised patch features, while the classification objective retains a gradient path to the representation. The local query control matches parameter capacity. The dictionary objective models patch reconstruction and sparsity; semantic object correspondence requires separate validation. The dictionary-without-penalties control obtains 55.05/70.93/39.17 mAP; adding penalties changes All/Known/Novel by +1.48/+2.50/+0.47 points. The observed penalty contrast includes execution-environment variation, which confounds attribution to the penalties.

C. The loader records the actual resize, crop, and horizontal flip. Inverse geometry maps overlapping patch centres between views. Nonoverlapping regions contribute neither positives nor negatives. The randomised control preserves the count of supervised positions while disrupting their correspondence. Given the potential global context in transformer patch features and the substantial crop overlap, geometric agreement is an imperfect proxy for label visibility. Relative to the global local-query reference, this component jointly varies spatial correspondence and the level of the discovery objective.

D. A full unlabelled 224-by-224 view is concatenated with a Known-only 224-by-224 donor to form a 224-by-448 input. The model processes the composed image in a separate forward pass. Only donor-positive Known entries receive additional positive supervision; donor-absent entries are unknown. The preservation arm adds Bernoulli KL between detached original-view Novel probabilities and the composed-view outputs. Composition preserves source pixels while altering the conditions for positional interpolation and attention. Their contributions to the Novel regression remain hypotheses requiring direct measurement.

Completed factorial contrasts. Training has been completed for all four A configurations. Variable minus fixed cardinality changes All/Known/Novel mAP by $-12.47/-19.82/-5.12$ points with uniform frequency, and $-15.07/-20.56/-9.58$ with relaxed frequency. Relaxed minus uniform frequency gives $+4.51/+4.29/+4.74$ with fixed cardinality and $+1.91/+3.54/+0.28$ with variable cardinality. Fixed/relaxed ran in a different execution environment from the other A cells. Execution-environment differences confound a causal interpretation of the factorial interaction.

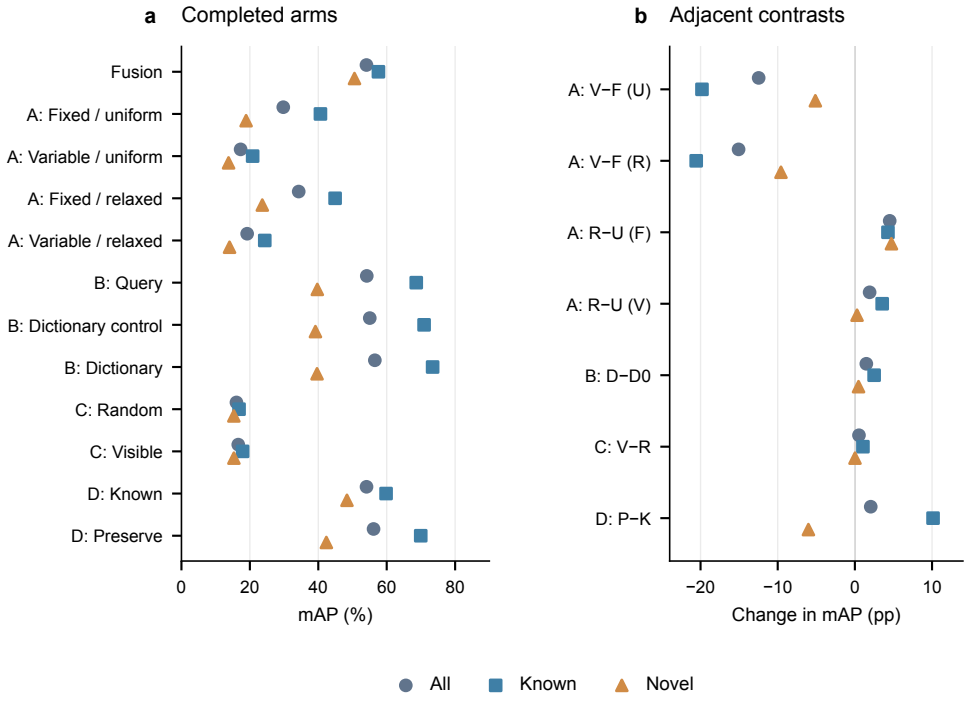


Figure 3: **Completed component study.** (a) All twelve full-horizon component arms, including the reference. (b) Within-design differences in percentage points. A labels vary cardinality under uniform/relaxed frequency (var–fix U/R) or relax frequency under fixed/variable cardinality (R–U fix/var). B compares dictionary penalties with the unpenalised dictionary; C compares visible with randomised correspondence; D adds Novel preservation. Single-seed values have no estimated error bars. A fixed/relaxed and the B unpenalised control cross execution environments relative to their counterparts.

F MIRFLICKR RANKING AND THRESHOLDED PREDICTION

The paired intervention removes Bernoulli distillation on observed Known entries of labelled images while preserving the original batch-by-class denominator. The pair matches the dataset, split, manifests, pretrained weights, code and environment, optimiser, augmentations, seed, final checkpoint selection, and evaluator; the changed parameter is the anchor-protection policy. It isolates the anchor-protection policy within this earlier training recipe.

Table 11: The MIR pair. mAP/F1 are percentages and MAE is in labels per image. Both Novel macro and micro F1 equal zero for both methods. Known macro/micro F1 are 0.0244/0.0213 for the control and 17.6001/17.5060 for anchor protection.

Method	All	Known	Novel	Macro F1	Micro F1	MAE
ASL-supervised	29.40	42.92	15.88	0.01	0.01	3.80
+ Anchor protection	23.01	31.72	14.31	8.80	9.26	3.61

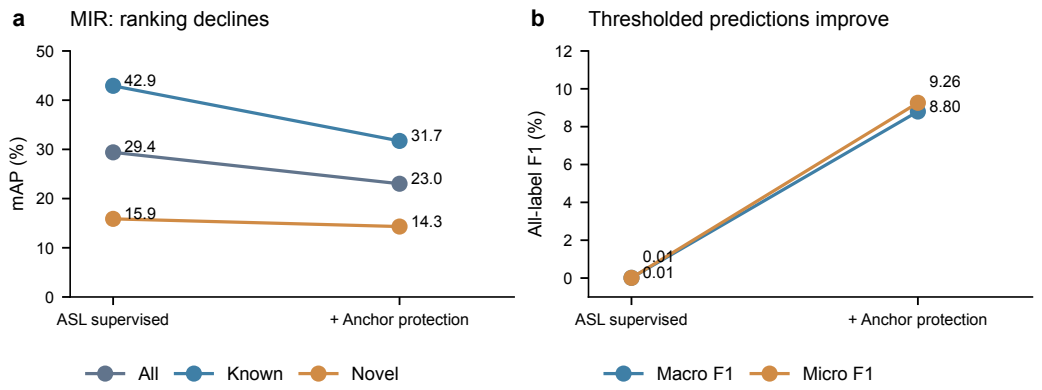


Figure 4: **An observed calibration–ranking disagreement on MIR.** The same full-horizon seed-0 pair lowers all three mAP components while increasing fixed-threshold All F1. Lines connect the two interventions evaluated on the same 4,917 held-out images. Both panels use the same threshold and matching rule.

G PAIRWISE NOVEL-SLOT AP

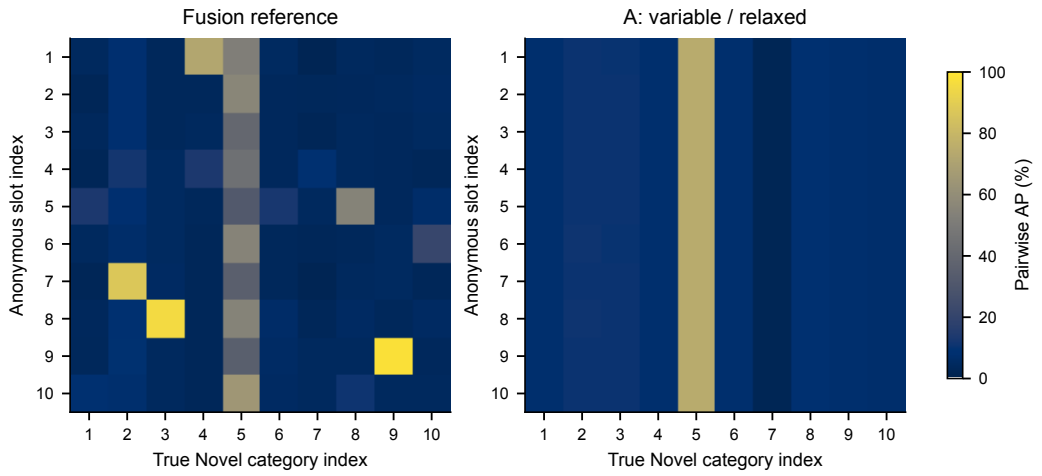


Figure 5: **Novel-slot differentiation in pairwise AP profiles.** Rows index anonymous outputs and columns index true Novel classes in fixed dataset order. The panels share a 0–100 scale. The variable/relaxed A arm has nearly indistinguishable row profiles, although its AP-maximising Hungarian permutation is well defined. This pattern indicates weak differentiation in evaluator-level profiles. Raw-vector identity and concept localisation require separate analyses. The matrices use all ten slots and all ten Novel labels on the same 4,952 test images; there is no subgroup rematching.

H EARLIER DISTILLATION CONFIGURATIONS AND THRESHOLDED METRICS

The six configurations form a cumulative ladder: Control injects no KD loss; Bernoulli adds frozen Known-head supervision; Relation also adds anchor and neighbourhood losses; Patch also adds aligned patch preservation; Dual also adds Novel EMA pseudo-labels; Adaptive changes their thresholds. Every arm retains the underlying sparse-discovery objective. The shared augmentation first resizes the complete image to 224×224 and applies a shared horizontal flip. The strong view adds ColorJitter with brightness/contrast/saturation/hue strengths $0.4/0.4/0.2/0.1$, and a nine-pixel Gaussian blur with $\sigma_{\text{blur}} \in [0.1, 2.0]$ with probability 0.5. Neither view uses cropping or eras-

Table 12: Historical seed-0 visual-transfer configurations, kept separate from the current four-worker three-seed panel. mAP and F1 are percentages; MAE is labels per image.

Method	All	Known	Novel	Macro F1	Micro F1	MAE
Control	56.43	70.95	41.91	36.98	38.61	2.19
Bernoulli	52.58	60.62	44.54	46.85	48.49	1.07
Relation	58.62	62.83	54.41	48.17	48.72	1.02
Patch	58.80	60.61	56.99	47.03	47.78	1.14
Dual	60.82	80.39	41.25	53.24	49.91	0.97
Adaptive	60.86	77.70	44.03	50.95	47.35	1.13

ing. This differs from the historical random-crop reference; the corresponding controls are therefore reported separately.

The Known teacher head is trained during the first 30 epochs using labelled valid Known entries and is then frozen. Its DINO backbone is always frozen. Known anchors normalise the accumulated teacher CLS features of valid positive labelled occurrences sampled during the first 30 epochs; repeated visits count separately. Bernoulli KL uses temperature one and weight one. Anchor, neighbour, and patch terms have weight 0.5 each. The teacher selects ten nearest neighbours within the minibatch, excluding self and repeated image identities; Huber regression matches their feature cosines without asserting common class membership. The patch term compares all 196 corresponding patch tokens. All KD contributions start at epoch 31 and ramp over ten epochs. This configuration implements fixed-teacher relation and visible-patch transfer with the losses and schedule specified above.

Dual maintains an EMA of student weights with cosine-scheduled momentum from 0.996 to 0.9999. On $\mathcal{D}_U$, weak-view probabilities above 0.8 are positive and below 0.2 are negative; other entries are ignored. Positive and negative BCE means are averaged over the branches that are present. Consequently, when a minibatch has no positive entries, the negative branch receives weight one instead of one half. This reduction increases the relative weight of negative-only batches. Adaptive uses per-slot moving 90th/10th probability quantiles with momentum 0.95, clipped to $[0.70, 0.95]$ for positives and $[0.02, 0.20]$ for negatives. Appendix M describes a fixed-branch variant evaluated at an intermediate checkpoint.

Table 13: Distillation F1 vector at threshold 0.5, in percent. A/K/N denote All/Known/Novel; all groups reuse the global Novel mapping. These six rows complement Table 12.

Method	A-macro	A-micro	K-macro	K-micro	N-macro	N-micro
Control	36.98	38.61	41.51	42.38	32.45	34.88
Bernoulli	46.85	48.49	54.66	57.81	39.05	42.07
Relation	48.17	48.72	54.51	56.73	41.82	42.94
Patch	47.03	47.78	50.53	51.85	43.54	44.41
Dual	53.24	49.91	69.00	73.41	37.49	40.09
Adaptive	50.95	47.35	67.79	72.52	34.12	37.53

I Evidence assignment and saved-prediction audit

Negative uses Known-only split membership to supervise Novel absence on labelled images. Balance changes per-category positive/negative normalisation. Multi combines these changes with full-image composition supervision. Unified instead uses local prototype/null competition to produce Novel evidence and replaces the earlier Novel-discovery/EMA supervision paths. Its student retains ten Novel slots. These variants are reported as exploratory development results guided by earlier observations.

Recorded audits cover the six distillation and four evidence-assignment arms, using each model's global test permutation for subgroup evaluation. The groups here are defined by *valid* annotations, giving 1,692 Known-only and 1,037 mixed test images. The validity mask potentially changes group membership for images whose only annotations in a group are marked difficult, relative to

Table 14: Completed evidence-assignment runs: mAP, All macro/micro F1 (percent), and cardinality MAE. All run for 200 epochs and 15,600 updates with seed 0.

Method	All	Known	Novel	Macro F1	Micro F1	MAE
Negative	59.82	81.29	38.35	49.07	45.57	1.24
Balance	59.87	79.66	40.07	51.42	49.56	0.89
Multi	56.76	76.62	36.90	48.40	44.80	1.28
Unified	51.38	83.27	19.49	37.48	40.74	1.65

Table 15: Full F1 vector for the evidence-assignment runs, in percent.

Method	A-macro	A-micro	K-macro	K-micro	N-macro	N-micro
Negative	49.07	45.57	64.25	71.35	33.88	36.96
Balance	51.42	49.56	65.92	71.80	36.93	40.53
Multi	48.40	44.80	64.05	70.60	32.75	36.09
Unified	37.48	40.74	57.72	67.91	17.25	30.74

an unmasked annotation-presence inventory. Images without valid positives are excluded from both positive groups.

Table 16: Composition audit under one global mapping per model. Novel FP counts are false-positive image–class entries on 1,692 valid-annotation Known-only test images; mixed recalls use 1,037 mixed images. Precision and recalls are percentages at threshold 0.5.

Method	Known-only Novel FP	Mixed K recall	Mixed N recall	Mixed N precision
Control	2303	82.09	36.20	26.44
Bernoulli	1349	44.82	30.97	27.72
Relation	1106	51.76	39.74	30.03
Patch	824	54.91	39.98	30.91
Dual	2019	41.31	42.32	27.06
Adaptive	2365	39.42	43.52	24.35
Negative	1722	29.07	47.87	20.94
Balance	1860	34.92	44.65	26.97
Multi	1845	30.42	48.59	20.76
Unified	153	27.45	48.27	22.36

Unified reduces false positives while showing incomplete category coverage on both the training and test sets. Its Novel mAP is 19.58 on the unlabelled training pool and 19.49 on test. It detects zero of 97 valid sheep positives and two of 259 valid train positives on test. Its mixed-image Novel recall exceeds Patch's, but precision is only 22.36%, illustrating the coexistence of aggregate recall gains and missing categories. Approximately 92.7% of positive evidence in the sampled training-log records belongs to two slots. At a strict pairwise prediction-correlation threshold of 0.99, Patch, Dual, and Unified have 0, 3, and 1 duplicate slot pairs. This statistic measures near-perfect pairwise correlation; shared low-dimensional variation and semantic duplication below the cutoff remain unassessed.

J A MATCHED BALANCE-GRADIENT INTERVENTION

A separate pair directs the marginal-balance loss either through the full category logits (Coupled) or only through category biases (Bias-only). Other losses and paired training/evaluation conditions are held fixed. Both runs finish 200 epochs. A first-step diagnostic reduces the balance-to-logit gradient norm from 0.00215648 to zero, while the balance-to-bias norm remains 0.02349617. This first-step diagnostic validates the intended gradient routing; the following performance comparison assesses its effect on discovery.

Bias-only routing lowers Novel mAP by 5.38 points while improving cardinality MAE from 2.26 to 2.10. Known mAP changes by only +0.21 points. Bias-only routing improves category-count accuracy while reducing Novel ranking performance.

Table 17: Balance-gradient routing: mAP, All macro/micro F1 (percent), and cardinality MAE.

Method	All	Known	Novel	Macro F1	Micro F1	MAE
Coupled balance	50.04	51.68	48.39	35.36	36.24	2.26
Bias-only balance	47.45	51.90	43.01	34.47	36.38	2.10

Table 18: Balance-gradient routing F1 vector, in percent.

Method	A-macro	A-micro	K-macro	K-micro	N-macro	N-micro
Coupled balance	35.36	36.24	41.57	42.82	29.15	30.11
Bias-only balance	34.47	36.38	41.05	43.07	27.89	30.08

K FROZEN-FEATURE EXTERNAL ADAPTATIONS

DLCL-Mask (Huang et al., 2020), MCUL-Mask/MCUL-Zero (Huang et al., 2021), and a semi-NMF plus ridge predictor use the same VOC class partition and evaluation convention, with DINO-v1 ViT-B/16 CLS features fully frozen. Images are resized to 224 pixels, features are ℓ_2 normalised, and a min–max transform is fitted on training features only. Each optimisation runs for 2,000 full-matrix iterations. This full-matrix optimisation schedule is distinct from 200 epochs of visual-network training. The panel uses fully frozen representations, separately from the fine-tuned image-training models.

Table 19: Frozen-feature adaptations, reported separately from fine-tuned models. Values are mAP and All macro/micro F1 (percent), plus cardinality MAE. The rows use the feature and solver adaptations specified below.

Method	All	Known	Novel	Macro F1	Micro F1	MAE
DLCL-Mask adaptation	40.93	72.66	9.20	33.11	33.63	1.11
MCUL-Mask adaptation	52.00	81.14	22.85	0.00	0.00	1.42
MCUL-Zero adaptation	51.84	81.08	22.61	0.00	0.00	1.42
Semi-NMF + ridge	41.30	71.50	11.09	36.81	34.66	1.21

These are formula-level adaptations with implementation differences. The DLCL solver uses ordinary proximal gradient rather than the source algorithm's accelerated FISTA update. MCUL uses projected backtracking for its latent matrix and a dimensional-consistency amendment to the implemented expression. Mask and Zero differ in whether unobserved Known entries are masked or filled with zero. The semi-NMF plus ridge predictor provides a simple factorisation-based reference. Newer DMNL and other adaptations appear separately in Appendix B.2.

Independent local recomputation from all four saved prediction arrays agrees with stored AP, global assignment, F1, and cardinality errors to absolute error below 10^{-7}. The file field named `probabilities` stores raw, uncalibrated linear scores for these models. MCUL-Mask and MCUL-Zero have maximum scores approximately 0.3408 and 0.3384: threshold 0.5 therefore predicts no positives, giving exact zero F1 despite nonzero ranking AP. The MCUL kernel has mean off-diagonal similarity approximately 1.52×10^{-5}. DLCL and semi-NMF also retain substantial terminal objective variation (approximately 18.9% over the last twenty recorded values); convergence has not been established. Score calibration and solver convergence are therefore additional factors in interpreting this panel.

Table 20: Frozen-feature adaptation F1 vector, in percent, at the common threshold 0.5.

Method	A-macro	A-micro	K-macro	K-micro	N-macro	N-micro
DLCL-Mask adaptation	33.11	33.63	66.01	67.49	0.21	0.24
MCUL-Mask adaptation	0.00	0.00	0.00	0.00	0.00	0.00
MCUL-Zero adaptation	0.00	0.00	0.00	0.00	0.00	0.00
Semi-NMF + ridge	36.81	34.66	68.94	68.48	4.68	6.38

L TRAINING-ONLY CATEGORY-COUNT ESTIMATION

These experiments use training images and permitted Known supervision to fit and select local factors. Known-class hiding creates proxy tasks: selected Known identities are withheld from fitting, then used only for an audit of count and assignment. True Novel labels and test images are not used. The blind estimate counts local factors with unresolved semantic identity; potential contents include background, object parts, and remaining Known concepts. The VOC recognition experiments accompanying these count studies use ten supplied Novel slots.

Earlier candidate families. Four initial full-training-feature studies compare an offline dictionary (A), global gating (B), split/merge selection (C), and a Beta–Bernoulli factor model (D). Table 21 reports the full small proxy suite, including the zero-hidden-class case. These experiments fit factors offline on fixed visual features. Family D returns a nonzero count on the null episode, indicating overestimation when no class is held out; its nonzero proxy AP coexists with incomplete recovery of held-out classes.

Table 21: Initial counting families. The proxy tasks hide 0, 3, and 3 Known classes; the two nonzero episodes use different class subsets. AP is zero-padded for missing true columns, in percent. Blind counts are training-side factor estimates; AP evaluates held-out Known categories.

Family	Blind	Null count	Hide-3a count	Hide-3b count	AP-3a	AP-3b
A	0	4	4	0	18.61	0.00
B	0	0	0	0	0.00	0.00
C	0	0	3	0	16.21	0.00
D	4	4	4	4	30.73	31.13

Four-arm fitting/selection study. A later reference, denoted V2, uses a selected candidate count of eight. The current Count V3 study crosses its original fitting/selection choices with weighted null-aware fitting (Fit-only), a revised region-average utility (Utility-only), and both (Joint). All four runs have completed thirteen Known-hidden episodes. Fit, selection, and audit use separate training-side image subsets; the new held-out proxy audit is distinct from the development proxy episodes. This audit evaluates transfer to held-out Known categories.

Table 22: Completed Count V3 study. Candidate count is the blind training-side estimate. Development and new-audit MAE measure errors on Known-hidden proxy tasks; audit AP is averaged over nonzero-hidden episodes, and Null FP is the any-slot false-positive rate on the null audit (both percent).

Method	Candidate count	Dev. MAE	New audit MAE	Audit AP	Null FP
V2	8	1.29	0.50	54.47	2.87
Fit-only	3	0.57	1.00	46.11	0.23
Utility-only	0	1.29	1.00	38.96	1.32
Joint	0	1.57	1.67	0.00	0.00

The selected capacities are 16, 4, 0, and 0 for V2, Fit-only, Utility-only, and Joint. Two image-half refits assess sample sensitivity, yielding blind-count ranges [8,8], [2,3], [0,0], and [0,0]. Null-episode count estimates are 0, 1, 0, and 0. On the new proxy audit, V2 attains lower count MAE than all three revisions, while both utility-changing arms select empty sets.

At capacity 16, weighted fitting increases prototype self-acceptance from 4/16 to 16/16. Joint retains eleven candidate slots supported by at least two distinct images, but the selection objective deletes

every slot. At capacity four its four candidates have image supports [2,12,17,88]; deletion changes the objective from -0.02535 to the empty set's zero. The capacity-16 objective similarly rises from -0.06766 to zero by deleting all slots. Replacing image-max evidence with region-average evidence while retaining a per-class cost of 0.005 changes their relative scale; normalising view disagreement by the positive union while retaining weight 0.25 is another coupled change. These coupled scale changes are candidate contributors to empty-set selection. Their individual effects remain unresolved and require term-wise analysis with fixed support and reassignment after deletion.

M REGIONAL EVIDENCE VARIANT: INTERMEDIATE-CHECKPOINT ANALYSIS

This regional-student variant is separate from PaD and from the category-count study. The available comparison is an epoch-60 training-side audit; final student test results are unavailable in the retained record.

Same-epoch interim observation. Table 23 compares epoch 60 (4,680 updates) on all 4,175 unlabelled *training* images. Their annotations are opened only for post-hoc audit and do not enter optimisation. These results are separate from the 37 completed performance rows. Hash/configuration checks establish two one-factor comparisons: U-BCE versus U0 changes only the fixed-BCE switch; U-Local versus U-Local0 changes only the region-loss weight. Both comparisons hold other recorded configuration factors fixed and use the same intermediate checkpoint.

Table 23: **Training-side audit at the intermediate checkpoint.** All arms use epoch 60, seed 0, 4,175 training images, and threshold 0.5. mAP/F1 are percentages and MAE is labels per image. These intermediate training-set diagnostics are separate from final-checkpoint test evaluation.

Method	All mAP	K mAP	N mAP	A micro F1	N micro F1	MAE
U0	50.31	78.45	22.16	29.12	22.65	2.08
U-BCE	51.11	78.69	23.54	29.29	23.26	2.29
U-Local0	54.42	77.93	30.92	37.00	23.47	1.04
U-Local	55.58	78.30	32.85	41.56	31.70	0.98

Fixed BCE raises Novel mAP by 1.38 points but increases cardinality MAE. Adding the local loss within the same regional architecture raises Novel mAP by 1.93 points and Novel micro F1 by 8.22 points, while Known micro F1 slightly declines. The horse class accounts for 71.5% of the net Novel mAP increase; U-Local's sheep, train, and sofa AP remain 2.93%, 4.96%, and 6.99%. These category-wise AP values characterise ranking; recall requires an explicit thresholded evaluation.

Both regional arms have Novel mAP 48.21% at epoch 20; their epoch-40 values are 30.87% and 32.08%, and epoch-60 values 30.92% and 32.85%. The identity switch occurs before epoch 31, at step 2,340, together with KD activation and replacement of the discovery loss. The observed transition coincides with simultaneous head, KD, and discovery-loss changes, leaving their individual contributions unresolved. The frozen protocol specifies the final checkpoint for evaluation rather than the epoch-20 score. Recent U-Local logs contain approximately 5,980 positive events, with 68.0% in the top two slots and one slot with none. These counts aggregate positive events without deduplicating image identities; distinct-image coverage remains unmeasured. Final-test recovery and category coverage remain unestablished.

Fixed positive/negative branch weights. The regional V3 correction gives each Novel class fixed positive and negative weights, with absent branches contributing zero. For Novel slot r, let $\mathcal{P}_r^+$ and $\mathcal{P}_r^-$ be the instance-index sets selected for positive and negative pseudo-supervision. Its student logit is $z_{i,K+r}$. The reduction is

$$\mathcal{L}_{\text{fixed}} = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{\frac{1}{2}\sum_{i\in\mathcal{P}_r^+}\operatorname{softplus}(-z_{i,K+r})}{\max(1,|\mathcal{P}_r^+|)} + \frac{\frac{1}{2}\sum_{i\in\mathcal{P}_r^-}\operatorname{softplus}(z_{i,K+r})}{\max(1,|\mathcal{P}_r^-|)}\right]. \tag{19}$$

An absent positive branch no longer doubles the negative coefficient. This differs both from averaging the available branches and from averaging only active categories.

Shared regional evidence. Weighted spherical fitting retains null competition at each iteration and uses region weights in the prototype updates; zero-mass slots are not forcibly populated. Region construction pools normalised DINO patch features into 1×1, 2×2, and 4×4 cells (21 regions), or a separate 4×4-only mode (16 regions). The local teacher, regional student, and inference path use the same region definition, with same-position cross-view agreement. At epoch 31 the Novel-head identity handoff resets only the Novel rows and their optimiser state. Novel inference takes the second-largest regional response *within each category*; category outputs are thresholded independently. U-Local0 and U-Local share the regional architecture and differ in local-loss weight (0 versus 0.25). Log-domain Sinkhorn is disabled for this batch.

Teacher support and acceptance. The fixed ten-slot teacher has per-slot cross-image supports [0,5,0,5,2,18,30,0,0,0]. Five slots have support from at least two images, seven prototype centres pass self-acceptance, and only three satisfy both checks. Positive evidence is present in 6.86% of selection images, and 98.65% of regions are rejected. A training-side Known-only audit gives any-slot false positives of 1.62%, passing the preregistered 10% implementation-admission bound. This gate evaluates false-positive control on Known-only images; semantic coverage remains unverified. Seven Known-hidden proxy episodes have matched AP 39.61% and count MAE 0.86. Their true/estimated counts are [0,1,1,2,2,3,3] and [1,1,1,1,3,5,2]. A separate blind candidate count of six is diagnostic only: the students retain ten Novel slots.

Fitting and acceptance remain imperfectly aligned: two supported prototypes reject their own centres. The deletion diagnostic masks previously assigned winners while retaining the original assignments. A complete leave-one-out utility calculation requires reassignment after deletion. Prototype self-acceptance and deletion utility remain unresolved; final-test validation of method-level benefits remains unavailable.

Relation to the main study. These exploratory count and regional variants examine different mechanisms from PaD's fixed-reference feature transfer. They are reported with their checkpoint, supervision, and configuration boundaries; their intermediate results are not included in the main method comparison.