FAMILIARITY-AWARE MIXTURE OF ENERGIES FOR MULTIMODAL GENERALIZED CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized category discovery (GCD) seeks to recognize labeled old classes while discovering unlabeled novel classes from mixed data. In the era of foundation models, a class that is absent from the labeled training set may still be partially known to a pretrained vision-language model. However, novel class names and labels remain unavailable during discovery, so the model must organize nameless mixed data into coherent categories. Moreover, novel categories are not equally familiar to foundation models, since some are well represented in the pretrained CLIP space whereas others remain long-tailed, fine-grained, or weakly grounded. Treating semantic evidence with a fixed or globally shared strategy can therefore lead to both semantic overtrust for unfamiliar clusters and semantic underuse for familiar ones. In this paper, we propose FAME, a Familiarity-Aware Mixture-of-Energies framework for multimodal GCD. FAME estimates cluster familiarity, an unsupervised reliability variable that measures whether CLIP image-space semantics provide stable and discriminative evidence for each nameless cluster, without using ground-truth novel class names or an external concept bank during discovery. It then performs risk-aware energy routing, adaptively combining CLIP-space semantic evidence, local visual cues, and self-supervised cluster structure for category assignment. This design strengthens semantic guidance for familiar and low-risk clusters while shifting uncertain or high-risk clusters toward visual-structural evidence. Assignments, prototypes, familiarity scores, and risk scores are jointly refined in an iterative process. Extensive experiments on standard GCD benchmarks demonstrate the effectiveness of FAME under semantically heterogeneous novel clusters.

1 INTRODUCTION

Generalized category discovery (GCD) studies how to recognize labeled old classes and discover unlabeled novel classes within the same data stream. It arises from the limitation of closed-set recognition, where a model assumes that all test samples belong to a fixed label space. In real open environments, new categories appear without manual annotation, and unlabeled data often contains both old and novel classes. GCD therefore provides a practical formulation for expanding the category space of a visual model while still using supervision from known categories.

The emergence of large foundation models changes the meaning of unknown categories in GCD. In the conventional setting, a class is novel because it is absent from the labeled training set. In the foundation-model era, this does not imply that the class is unknown to the model. Vision-language models such as CLIP are pretrained on large-scale image-text data and encode broad open-world semantics. Some benchmark novel categories are already well represented in the pretrained semantic space, while others remain long-tailed, fine-grained, weakly grounded, or rarely observed. Thus, novel categories are not uniformly unknown to foundation models; they exhibit different degrees of foundation-model familiarity.

Recent GCD methods make increasing use of foundation models. Some methods build on pretrained visual encoders and local visual cues Vaze et al. (2022a); Wen et al. (2023); Zhang et al. (2023b). Others use vision-language representations, text prompts, class descriptions, and generated semantic concepts Ouldnoughi et al. (2023); Zheng et al. (2024); Wang et al. (2025b). These methods show that semantic information from foundation models helps category discovery. Most methods, however,

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.

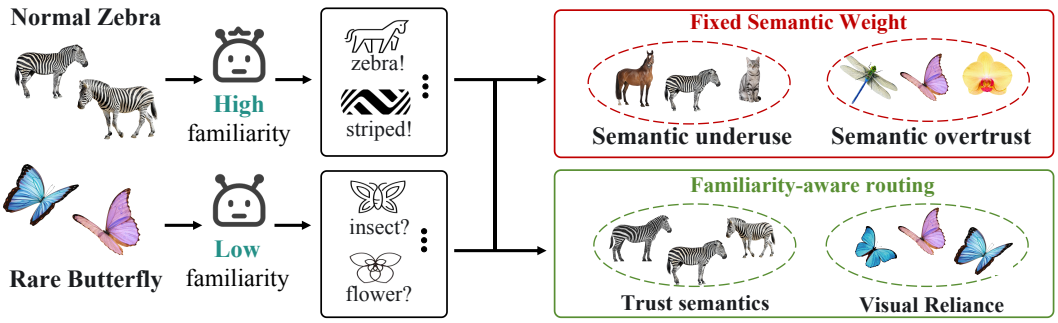

Figure 1: Fixed weighting vs. familiarity-aware routing.

introduce this information through fixed losses, shared prompts, or global fusion rules over unlabeled samples and clusters. They rarely examine whether the semantic signal has the same reliability across different novel clusters. This matters because foundation models represent some novel categories well, while they provide much weaker support for others.

Such heterogeneity makes a uniform semantic fusion rule unsuitable. A single global semantic weight does not fit all novel clusters. As illustrated in Fig. 1, this mismatch leads to two failure modes. **Semantic overtrust** occurs when a cluster receives weak support from the foundation model. Its semantic cues are noisy, biased toward old classes, or drawn toward frequent pretraining patterns. Giving these cues too much weight misleads category assignment and weakens the visual structure of novel clusters. **Semantic underuse** occurs when a cluster is well represented in the pretrained space. Its semantic cues provide reliable and discriminative evidence, especially when visual differences are subtle, but a fixed fusion rule does not make full use of this information. The key issue is not whether semantic evidence helps GCD, but when it is reliable. We define the *familiarity* of a nameless cluster as the extent to which a pretrained vision-language model provides stable and discriminative CLIP-space evidence, without querying novel class names or using an external concept dictionary. As shown in Fig. 2, unlike old-versus-novel confidence or visual compactness, familiarity captures the reliability of semantic evidence in the pretrained CLIP space. High-familiarity clusters have stable and well-separated assignments, while low-familiarity clusters show unstable or ambiguous evidence and rely more on visual and structural cues.

We therefore propose **FAME**, a **F**amiliarity-**A**ware **M**ixture of **E**nergies framework for multimodal generalized category discovery. FAME estimates not only whether foundation-model semantics are reliable for a nameless cluster, but also whether using such semantics is safe for category discovery, without using ground-truth novel class names. It contains three connected components. The first component is **Counterfactual Familiarity Estimation**, which measures semantic reliability by testing whether CLIP-space assignments remain stable and semantically anchored under image- and feature-level perturbations. The second component is **Semantic Intervention Risk**, which estimates the potential harm of semantic guidance by measuring visual-semantic assignment conflict and CLIP-space ambiguity. The third component is **Risk-Aware Energy Routing**, which adaptively combines CLIP-based semantic evidence, local visual cues, and self-supervised cluster structure. In this routing process, familiar and low-risk clusters receive stronger semantic guidance, while uncertain or high-risk clusters rely more on visual-structural evidence. Assignments, prototypes, familiarity scores, and risk scores are refined iteratively.

2 Related Work

2.1 Generalized Category Discovery

Generalized category discovery (GCD) studies how to categorize unlabeled data containing both known and novel classes, given supervision from only a subset of known categories Vaze et al. (2022a). Existing methods typically learn transferable visual representations and perform clustering or prototype-based assignment, using contrastive learning Pu et al. (2023), prompt-based affinity modeling Zhang et al. (2023a), probabilistic modeling Zhao et al. (2023), parametric classification with category prototypes Wen et al. (2023), feature enhancement and representation sharpening, or

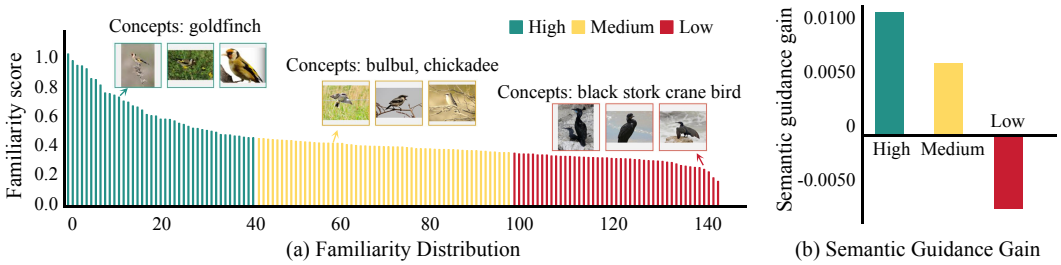

Figure 2: Semantic familiarity heterogeneity in nameless novel clusters. (a) Clusters are sorted by familiarity scores computed from CLIP-space stability and semantic anchoring, without using ground-truth novel class names. Displayed concepts are manually inspected only for post hoc visualization and are not used by FAME during discovery. (b) Semantic guidance helps high and medium familiarity clusters but hurts low familiarity clusters, supporting adaptive semantic routing.

preset feature geometry Xiao et al. (2024); Han et al. (2025b; 2026a). Recent methods further improve clustering reliability with contrastive mean-shift learning Choi et al. (2024) and begin to exploit CLIP-based language priors through prompt tuning or dual-branch vision-language learning Yang et al. (2025); Wang et al. (2025a). Despite these advances, most GCD methods still focus on visual discriminability and clustering reliability, while treating novel categories as relatively homogeneous during discovery. In practice, different nameless novel clusters can receive semantic evidence of very different quality from foundation models, especially in fine-grained scenarios where some clusters correspond to familiar visual concepts while others activate vague, generic, or background-related semantics. This work studies such semantic familiarity heterogeneity and uses it to guide adaptive semantic reasoning in GCD.

2.2 Vision-Language Semantic Guidance

Vision-language models learn aligned image-text representations from large-scale paired data and provide semantic priors beyond fixed label spaces Tan & Bansal (2019); Huang et al. (2021); Lu et al. (2019); Radford et al. (2021); Yao et al. (2021). In open-vocabulary and zero-shot recognition, class names, prompts, attributes, or descriptions are used as semantic anchors for visual recognition Zhou et al. (2022); Zhang et al. (2022); Han et al. (2025a); Dong et al. (2025); Peng et al. (2025); Wang et al. (2026). Recent GCD methods also introduce CLIP features, textual prompts, class-name embeddings, or image-text alignment objectives to improve novel category discovery, showing that vision-language priors can help connect unlabeled visual groups with semantic concepts. However, novel class names are unavailable during GCD, so semantic guidance must be inferred from nameless clusters rather than queried directly. This makes the reliability of semantic evidence cluster-dependent, because a pretrained vision-language model may provide confident cues for common categories but vague or biased cues for rare and fine-grained ones. Most existing vision-language guided methods use semantic priors relatively uniformly, implicitly assuming similar reliability across unlabeled samples or clusters. This assumption is unstable in fine-grained GCD: some novel clusters align with clear open-vocabulary concepts, while others only match generic or background semantics. Our work therefore estimates the reliability of semantic evidence for nameless clusters and uses it to adapt semantic guidance.

3 Problem Formulation

Let $\mathcal{D}_l = \{(x_i, y_i)\}$ denote the labeled set from old classes $\mathcal{Y}_o$, and let $\mathcal{D}_u = \{x_i\}_{i=1}^{N}$ denote the unlabeled set. The unlabeled set contains samples from both old and novel classes. In the known-K setting, the number of novel categories K_n is given. We maintain $C = |\mathcal{Y}_o| + K_n$ category slots, where the first $|\mathcal{Y}_o|$ slots correspond to old classes and the remaining K_n slots correspond to novel clusters. In the unknown-K setting used in our experiments, K_n is first estimated following the same category-number estimation protocol as the compared GCD baseline, and FAME then operates on the resulting estimated slots. This keeps the routing mechanism independent of ground-truth novel class counts. We use a visual encoder $f_v(\cdot)$ to extract normalized visual features $\mathbf{z}_i = \mathrm{norm}(f_v(x_i))$. Old class prototypes are initialized from labeled samples, while novel prototypes are initialized by

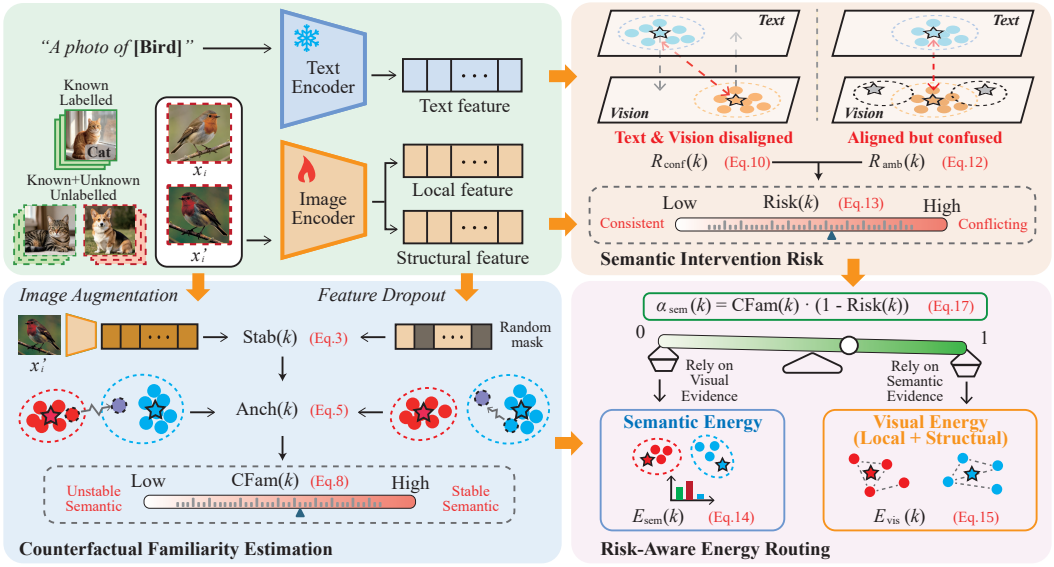

Figure 3: Overview of the proposed FAME framework. FAME estimates counterfactual familiarity from CLIP-space stability and semantic anchoring, estimates semantic intervention risk from visual-semantic conflict and CLIP-space ambiguity, and then performs risk-aware energy routing between semantic energy and visual-structural energy. Familiar and low-risk clusters receive stronger semantic guidance, while uncertain or high-risk clusters rely more on visual evidence.

a warm-up GCD model or clustering algorithm on unlabeled visual features. The goal is to assign each unlabeled sample to one of the C slots while preserving old-class recognition and discovering coherent novel categories.

4 METHOD

We propose **FAME**, a **F**amiliarity-**A**ware **M**ixture of **E**nergies framework for multimodal generalized category discovery. FAME is based on the observation that a category can be novel to the labeled training set but familiar to a foundation model. Some nameless clusters receive stable and discriminative evidence in the pretrained CLIP space, while others have unstable or ambiguous CLIP-space assignments. Therefore, semantic guidance should be used adaptively rather than uniformly. FAME follows a lightweight reliability-and-risk principle. As shown in Fig. 3. First, we estimate **counterfactual familiarity**, which measures whether the CLIP-space assignment of a cluster remains stable and semantically anchored under image- and feature-level perturbations. Second, we estimate **semantic intervention risk**, which measures whether CLIP-based semantic guidance may distort the visual cluster structure or remain ambiguous across competing category slots. Finally, we perform **risk-aware energy routing**, where CLIP-based semantic energy is emphasized only when the cluster is both familiar and low-risk, while high-risk clusters rely more on visual-structural energy.

4.1 COUNTERFACTUAL FAMILIARITY ESTIMATION

We estimate cluster familiarity without using an external concept bank or ground-truth novel class names. A cluster is considered familiar to the vision-language model when its CLIP-space assignment is stable under perturbations and clearly separated from competing slots. Let $\mathbf{u}_i = \mathrm{norm}(f^{\mathrm{ref}}_{\mathrm{clip}}(x_i))$ be the image feature from a frozen reference CLIP encoder. For each slot c, we maintain a CLIP-space prototype $\mathbf{p}^u_c$. The CLIP assignment energy is

$$E_{\mathrm{clip}}(\mathbf{u}_i, c) = -\mathbf{u}_i^\top \mathbf{p}^u_c. \tag{1}$$

We use two lightweight counterfactual interventions. The first applies weak image augmentation, including random resized crop and horizontal flip, and then re-extracts the CLIP feature. The second randomly masks a ratio ρ of CLIP feature dimensions before normalization. The perturbed feature is

denoted as $\mathbf{u}_i^b$, where $b \in \{\text{aug}, \text{drop}\}$. Its counterfactual assignment is

$$a_i^b = \arg\min_c E_{\text{clip}}(\mathbf{u}_i^b, c). \tag{2}$$

Let a_i denote the current slot assignment of x_i. For cluster C_k, counterfactual stability is defined as

$$\text{Stab}(k) = \frac{1}{2|C_k|} \sum_{x_i \in C_k} \sum_{b \in \{\text{aug},\text{drop}\}} \mathbb{I}[a_i^b = a_i]. \tag{3}$$

This score measures whether CLIP-space evidence continues to support the current assignment after image-level and feature-level perturbations. Stability does not capture how discriminative the assignment is. We therefore compute a semantic anchoring margin:

$$\gamma_i = \max\left(\min_{c \neq a_i} E_{\text{clip}}(\mathbf{u}_i, c) - E_{\text{clip}}(\mathbf{u}_i, a_i), 0\right). \tag{4}$$

The cluster-level anchoring score is

$$\text{Anch}(k) = \frac{1}{|C_k|} \sum_{x_i \in C_k} \gamma_i. \tag{5}$$

To avoid sensitivity to empty clusters or degenerate score ranges, we use a robust normalization over the set of non-empty clusters $\mathcal{K}_{\text{ne}}$. For a cluster-level score $s(k)$, let

$$s_{\min} = \min_{j \in \mathcal{K}_{\text{ne}}} s(j), \quad s_{\max} = \max_{j \in \mathcal{K}_{\text{ne}}} s(j). \tag{6}$$

We define

$$\text{Norm}(s(k)) = \text{clip}\left(\frac{s(k) - s_{\min}}{s_{\max} - s_{\min} + \epsilon}, 0, 1\right). \tag{7}$$

If $s_{\max} = s_{\min}$, we set $\text{Norm}(s(k)) = 0.5$ for all non-empty clusters. Empty clusters keep their previous EMA-smoothed scores. The familiarity score is then defined as

$$\text{CFam}(k) = \frac{1}{2}\left(\text{Norm}(\text{Stab}(k)) + \text{Norm}(\text{Anch}(k))\right). \tag{8}$$

Thus, $\text{CFam}(k)$ measures whether a cluster receives CLIP-space evidence that is both stable and discriminative.

4.2 Semantic Intervention Risk

Stable semantic evidence is not always useful for assignment. It can conflict with the visual structure learned by the GCD backbone, or it can remain ambiguous across similar slots. We estimate this potential cost with semantic intervention risk.

The first risk factor is visual-semantic conflict. We compare the assignment preferred by visual prototypes with the assignment preferred by CLIP prototypes:

$$a_i^v = \arg\max_c \mathbf{z}_i^\top \mathbf{p}_c^v, \quad a_i^u = \arg\max_c \mathbf{u}_i^\top \mathbf{p}_c^u. \tag{9}$$

Here $\mathbf{z}_i$ is the visual feature from the GCD backbone, and $\mathbf{p}_c^v$ is the corresponding visual prototype. The conflict risk is

$$R_{\text{conf}}(k) = \frac{1}{|C_k|} \sum_{x_i \in C_k} \mathbb{I}[a_i^u \neq a_i^v]. \tag{10}$$

The second risk factor is semantic ambiguity. We compute the CLIP-space assignment distribution with a fixed temperature τ_u:

$$s_{i,c} = \frac{\exp(\mathbf{u}_i^\top \mathbf{p}_c^u / \tau_u)}{\sum_{c'=1}^{C} \exp(\mathbf{u}_i^\top \mathbf{p}_{c'}^u / \tau_u)}. \tag{11}$$

The ambiguity risk is the normalized entropy:

$$R_{\text{amb}}(k) = \frac{1}{|C_k|} \sum_{x_i \in C_k} \frac{-\sum_{c=1}^{C} s_{i,c} \log s_{i,c}}{\log C}. \tag{12}$$

Using the robust normalization operator $\mathrm{Norm}(\cdot)$ defined in Sec. 4.1, the final risk score is

$$\mathrm{Risk}(k) = \frac{1}{2}\left(\mathrm{Norm}(R_{\mathrm{conf}}(k)) + \mathrm{Norm}(R_{\mathrm{amb}}(k))\right). \tag{13}$$

A higher risk score indicates stronger visual-semantic conflict or weaker CLIP-space separation.

4.3 Risk-Aware Energy Routing

FAME uses the estimated cluster-level reliability signals to route between CLIP-space semantic evidence and visual-structural evidence. The semantic energy is computed in the frozen CLIP image space:

$$E_{\mathrm{sem}}(x_i, c) = -\mathbf{u}_i^\top \mathbf{p}_c^u. \tag{14}$$

This term uses CLIP image-space features and prototypes, and does not query novel class-name prompts or external concept banks.

For visual evidence, we combine local and structural cues. The local feature $\mathbf{l}_i$ is obtained by average pooling the last-block patch tokens of the trainable image encoder, followed by ℓ_2 normalization. The structural feature $\mathbf{z}_i$ is the GCD visual representation. Their energies are combined as

$$E_{\mathrm{vis}}(x_i, c) = \frac{1}{2}\left(-\mathbf{l}_i^\top \mathbf{p}_c^l - \mathbf{z}_i^\top \mathbf{p}_c^v\right). \tag{15}$$

All features and prototypes are ℓ_2 normalized before energy computation, so both energy terms are cosine-based.

After warm-up GCD initialization, FAME iteratively updates assignments, prototypes, familiarity scores, and risk scores. The normalized familiarity and risk scores are smoothed with EMA:

$$\bar{s}^{(t)}(k) = m_s \bar{s}^{(t-1)}(k) + (1 - m_s) s^{(t)}(k), \tag{16}$$

where s denotes either CFam or Risk, and m_s is the score momentum. If a cluster is empty at iteration t, we keep $\bar{s}^{(t)}(k) = \bar{s}^{(t-1)}(k)$. Prototypes in the CLIP, local, and structural spaces are also updated with EMA.

For sample x_i, let $k(i)$ denote its current cluster assignment. We define the semantic routing coefficient as

$$\alpha_{\mathrm{sem}}(i) = \eta_t \cdot \mathrm{sg}\left[\overline{\mathrm{CFam}}(k(i))\left(1 - \overline{\mathrm{Risk}}(k(i))\right)\right], \tag{17}$$

where $\eta_t \in [0, 1]$ is a routing ramp-up coefficient and $\mathrm{sg}[\cdot]$ denotes stop-gradient. Thus, familiarity and risk are treated as EMA-smoothed cluster-level statistics rather than parameters optimized by direct back-propagation. The visual routing coefficient is

$$\alpha_{\mathrm{vis}}(i) = 1 - \alpha_{\mathrm{sem}}(i). \tag{18}$$

The final assignment energy is

$$E_{\mathrm{FAME}}(x_i, c) = \alpha_{\mathrm{sem}}(i) E_{\mathrm{sem}}(x_i, c) + \alpha_{\mathrm{vis}}(i) E_{\mathrm{vis}}(x_i, c). \tag{19}$$

The assignment probability is computed by applying a softmax over negative energies:

$$P(c|x_i) = \frac{\exp(-E_{\mathrm{FAME}}(x_i, c))}{\sum_{c'=1}^{C} \exp(-E_{\mathrm{FAME}}(x_i, c'))}. \tag{20}$$

The trainable encoder is optimized with the base GCD objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{GCD}} + \lambda_{\mathrm{bal}} \mathcal{L}_{\mathrm{bal}}. \tag{21}$$

Here $\mathcal{L}_{\mathrm{GCD}}$ is the base GCD loss, and $\mathcal{L}_{\mathrm{bal}}$ avoids degenerate cluster usage Wen et al. (2023). FAME therefore changes the assignment energy through non-parametric reliability-guided routing, rather than introducing an additional differentiable routing loss.

Table 1: Comparison with the state of the art on GCD. The best results are in bold.

	Method	CUB			Stanford Cars			FGVC Aircraft			CIFAR100			ImageNet100			Herbarium19		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
(a) Clustering with the ground-truth number of categories K given																			
DINO	GCD (CVPR 22)	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	73.0	76.2	66.5	74.1	89.8	66.3	35.4	51.0	27.0
DINO	GPC (ICCV 23)	55.4	58.2	53.1	42.8	59.2	32.8	46.3	42.5	47.9	77.9	85.0	63.0	76.9	94.3	71.0	-	-	-
DINO	SimGCD (ICCV 23)	60.3	65.6	57.7	53.8	71.9	45.0	54.2	59.1	51.8	80.1	81.2	77.8	83.0	93.1	77.9	44.0	58.0	36.4
DINO	CMS (CVPR 24)	68.2	76.5	64.0	56.9	76.1	47.6	56.0	63.4	52.3	82.3	85.7	75.5	84.7	95.6	79.2	36.4	54.9	26.4
DINO	SelEx (ECCV 24)	73.6	75.3	72.8	58.5	75.6	50.3	57.1	64.7	53.3	82.3	85.3	76.3	83.1	93.6	77.8	39.6	54.9	31.3
DINO	RLCD (ICML 25)	70.0	79.1	65.4	64.9	79.3	58.0	60.6	62.2	59.8	83.4	84.2	81.9	86.9	94.2	83.2	46.4	61.2	38.4
CLIP	GCD-CLIP	51.1	56.2	48.6	62.5	73.9	57.0	41.2	43.0	40.2	**84.2**	83.1	**82.3**	79.3	94.6	71.1	39.7	58.0	29.9
CLIP	CPT (IJCV 25)	70.1	73.5	68.4	74.2	84.3	69.3	53.3	56.1	51.9	81.3	81.3	81.2	89.2	95.2	86.1	45.3	61.2	36.7
CLIP	GET (CVPR 25)	77.0	78.1	76.4	78.5	86.8	74.5	58.9	59.6	58.5	82.1	85.5	75.5	91.7	95.7	89.7	49.7	64.5	41.7
CLIP	CMS-CLIP	65.8	75.3	61.1	77.9	89.0	72.6	50.3	59.1	45.9	80.3	85.2	70.6	85.9	93.8	82.0	36.2	56.5	25.3
CLIP	SimGCD-CLIP	69.6	75.8	66.5	69.4	82.2	62.9	53.5	57.9	51.7	81.1	84.2	75.2	89.9	94.2	87.5	47.9	62.3	39.9
CLIP	SelEx-CLIP	74.2	69.5	76.5	68.8	68.4	69.0	56.7	58.2	56.0	77.2	77.0	77.6	91.2	93.6	90.0	44.4	57.1	37.6
CLIP	AL-GCD (CVPR 26)	84.1	79.7	86.3	79.0	90.8	73.6	66.6	66.1	66.8	80.4	85.6	69.8	92.7	96.2	91.1	47.7	59.0	41.6
CLIP	**FAME**	**86.8**	**79.9**	**90.4**	**81.5**	**92.7**	**76.2**	**69.3**	**66.3**	**71.3**	82.5	**86.0**	76.1	**94.0**	**97.6**	**92.0**	**50.4**	**65.1**	**42.0**
(b) Clustering without the ground-truth number of categories K given																			
DINO	GCD (CVPR 22)	51.1	56.4	48.4	39.1	58.6	29.7	-	-	-	70.8	77.6	57.0	77.9	91.1	71.3	37.2	51.7	29.4
DINO	GPC (ICCV 23)	52.0	55.5	47.5	38.2	58.9	27.4	43.3	40.7	44.8	75.4	84.6	60.1	75.3	93.4	66.7	36.5	51.7	27.9
DINO	CMS (CVPR 24)	64.4	68.2	62.2	51.7	68.9	43.4	55.2	60.6	52.4	79.6	83.2	72.3	81.3	95.6	74.2	37.4	56.5	27.1
CLIP	CMS-CLIP (CVPR 24)	65.6	74.0	61.3	77.2	87.3	72.3	50.6	52.8	49.5	78.0	81.2	71.5	84.8	93.8	80.2	38.8	57.7	28.6
CLIP	AL-GCD (CVPR 26)	73.7	79.1	71.0	79.9	90.5	74.7	55.3	58.9	53.5	83.3	84.9	80.2	85.7	96.5	80.3	41.2	59.0	31.6
CLIP	**FAME**	**85.1**	**81.3**	**86.9**	**81.3**	**94.5**	**75.9**	**68.2**	**64.7**	**70.0**	**84.4**	**85.9**	**82.6**	**87.8**	**97.3**	**84.2**	**49.5**	**64.6**	**41.8**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. Our evaluation spans six image classification benchmarks, including both generic and fine-grained datasets. We use CIFAR-100 Krizhevsky (2009) and ImageNet-100 Deng et al. (2009) for generic datasets. For fine-grained datasets, we evaluate on CUB-200 Wah et al. (2011), Stanford Cars Krause et al. (2013), FGVC Aircraft Maji et al. (2013), and Herbarium19 Tan et al. (2019). To separate categories into known and unknown categories, we follow the SSB split protocol Vaze et al. (2022b) for the fine-grained datasets. For CIFAR-100 and ImageNet-100, we perform a random category split using a fixed seed, consistent with previous studies.

Implementation Details. We use CLIP ViT-B/16 Radford et al. (2021) as the backbone. For familiarity and risk estimation, we keep a reference CLIP image encoder frozen. The trainable recognition branch shares the CLIP initialization but freezes most parameters and updates only the last transformer block of the image encoder, together with a lightweight adapter and the projection head. The text encoder is kept frozen and is used only to encode labeled old-class names for old-class textual prototypes. We use the same hyperparameters for all datasets: $\rho = 0.2$, $\tau_u = 0.07$, prototype momentum $m = 0.9$, score momentum $m_s = 0.9$, base learning rate 0.1, adapter learning rate 1×10^{-3}, batch size 256, weight decay 1×10^{-4}, and two augmented views.

Comparison with State of the Art Methods. We compare FAME with recent GCD methods Vaze et al. (2022c); Zhao et al. (2023); Rastegar et al. (2024); Liu et al. (2025); Choi et al. (2024); Yang et al. (2025); Wang et al. (2025b); Han et al. (2026b). Following Choi et al. (2024), we report two settings: (a) clustering with the ground-truth number of categories and (b) clustering without it. In the latter setting, the number of novel categories is estimated before slot initialization using the same protocol as the corresponding baseline, and FAME uses only the estimated slots afterward. All reported results are averaged over three random seeds.

5.2 EXPERIMENTAL MAIN RESULTS

Tables 1 and 2 report the main results. FAME consistently improves over recent CLIP-based GCD methods on most datasets under both known-K and unknown-K settings. In the unknown-K setting, FAME follows the same category-number estimation protocol as AL-GCD, while each method uses its own estimated number of slots; more details and estimated category numbers are provided in the appendix. This setting is challenging because over-estimation may split one class into multiple

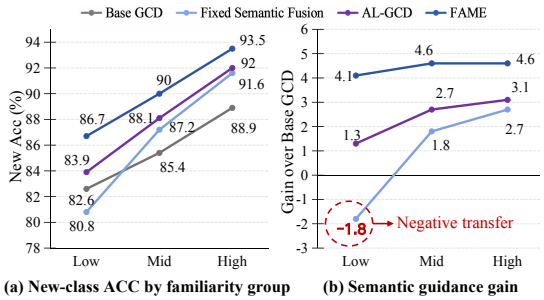

Figure 4: Familiarity-stratified evaluation on CUB.

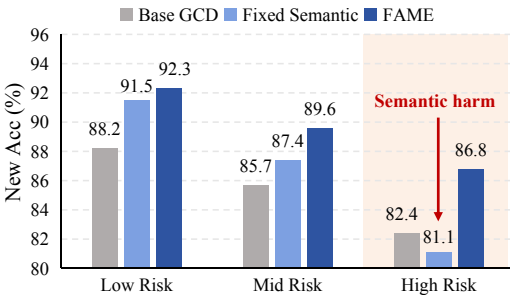

Figure 5: Risk validation on CUB.

Table 2: Comparison on Fine-grained Avg, Classification Avg, and All Datasets Avg with CLIP backbone.

Method	Fine-grained			Classification			All Datasets		
	All	Old	New	All	Old	New	All	Old	New
(a) Clustering with the GT number of categories K given									
CPT	65.9	71.3	63.2	85.3	88.3	83.7	68.9	75.3	65.6
GET	71.5	74.8	69.8	86.9	90.6	82.6	73.0	78.4	69.4
CMS	64.7	74.5	59.9	83.1	89.5	76.3	66.1	76.5	59.6
SimGCD	64.2	71.9	60.4	85.5	89.2	81.4	68.6	76.1	64.0
SelEx	66.6	65.3	67.2	84.2	85.3	83.8	68.8	70.6	67.8
AL-GCD	76.6	78.9	75.6	86.6	90.9	80.5	75.1	79.6	71.5
FAME	**79.2**	**79.6**	**79.3**	**88.3**	**91.8**	**84.1**	**77.4**	**81.3**	**74.7**
(b) Clustering without the GT number of categories K given									
CMS	64.5	71.4	61.0	81.4	87.5	75.9	65.8	74.5	60.6
AL-GCD	69.6	76.2	66.4	84.5	90.7	80.2	69.8	78.2	65.2
FAME	**78.3**	**80.2**	**77.6**	**86.1**	**91.6**	**83.4**	**76.1**	**81.4**	**73.6**

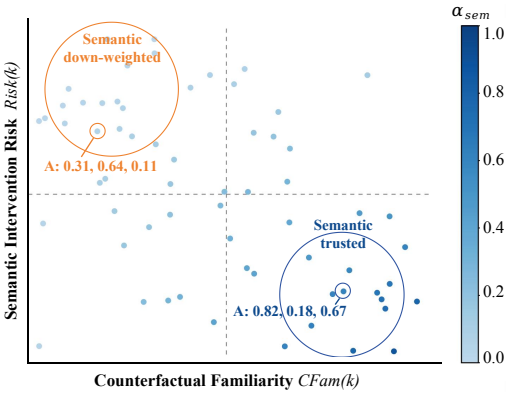

Figure 6: Cluster-level routing landscape of FAME.

clusters, while under-estimation may merge different classes. In the known-K setting, FAME achieves strong performance across both generic and fine-grained benchmarks. On ImageNet100, FAME improves All accuracy over AL-GCD from 92.7% to 94.0%. The gains are more pronounced on fine-grained datasets, where semantic evidence is more heterogeneous across novel clusters. FAME improves All accuracy over AL-GCD by 2.7, 2.5, and 2.7 points on CUB, Stanford Cars, and FGVC Aircraft, respectively, and improves Herbarium19 from 47.7% to 50.4%. The advantage of FAME becomes even clearer in the unknown-K setting. FAME improves All accuracy over AL-GCD from 85.7% to 87.8% on ImageNet100, from 73.7% to 85.1% on CUB, and from 55.3% to 68.2% on FGVC Aircraft. On Herbarium19, FAME improves All accuracy from 41.2% to 49.5%, further showing its robustness in long-tailed discovery scenarios.

5.3 Analysis of Semantic Reliability and Routing

Familiarity-stratified evaluation. We further examine the effect of semantic guidance under different levels of foundation-model familiarity. For this post hoc evaluation only, we use ground-truth novel labels to aggregate the cluster-level familiarity scores into a class-level score $\mathrm{Fam}_{cls}(y)$ on CUB, and then divide classes into low-, mid-, and high-familiarity groups. These labels are not used for training, routing, or model selection. Fig. 4 reports New Acc and the gain over Base GCD within each group. Fixed semantic fusion improves accuracy in mid- and high-familiarity groups, but its gain becomes negative in the low-familiarity group, showing that uniform semantic guidance can introduce negative transfer when semantic evidence is weak or generic. In contrast, FAME keeps positive gains across all groups and improves low-familiarity classes by reducing unreliable semantic influence.

Risk validation. Fig. 5 evaluates whether $\mathrm{Risk}(k)$ reflects harmful semantic intervention. After training, we sort novel clusters by $\mathrm{Risk}(k)$ and split them into three equal-sized groups, denoted as low-, mid-, and high-risk groups. New Acc is then reported within each group. Fixed semantic fusion improves performance in the low-risk group, where semantic concepts are more consistent with the visual structure. In the high-risk group, its accuracy drops below the base GCD backbone, indicating

Table 3: Comparison with simpler routing gates under the known-K setting. We report All, Old, and New clustering accuracy.

Routing strategy	CUB			Stanford Cars			FGVC Aircraft			CIFAR100			ImageNet100			Herbarium19		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Fixed semantic:visual	84.2	79.0	87.0	79.7	91.3	74.1	67.2	64.8	68.4	81.5	85.2	74.4	92.6	96.7	90.0	48.7	63.6	40.0
Random gate	84.4	79.1	87.2	79.8	91.4	74.2	67.3	64.9	68.5	81.6	85.3	74.5	92.7	96.8	90.1	48.8	63.7	40.1
Cluster-size gate	84.7	79.2	87.5	80.0	91.5	74.4	67.5	65.0	68.7	81.7	85.4	74.7	92.8	96.9	90.3	49.0	63.8	40.2
CLIP compactness gate	85.1	79.3	87.9	80.2	91.6	74.6	67.7	65.2	68.9	82.0	85.6	75.0	93.1	97.0	90.8	49.2	64.0	40.4
Visual compactness gate	84.9	79.3	87.7	80.1	91.5	74.5	67.6	65.1	68.8	81.8	85.5	74.8	92.9	96.9	90.5	49.1	63.9	40.3
Entropy-only gate	85.3	79.4	88.1	80.3	91.7	74.7	67.9	65.4	69.1	82.1	85.7	75.2	93.3	97.1	91.0	49.3	64.1	40.5
Margin-only gate	85.5	79.5	88.3	80.4	91.8	74.8	68.0	65.5	69.2	82.2	85.8	75.4	93.4	97.2	91.1	49.4	64.1	40.6
CFam-only routing	85.4	79.2	88.7	80.4	91.7	75.0	68.1	65.1	69.6	82.0	85.6	75.1	93.2	97.0	91.0	49.4	64.3	40.9
Risk-only routing	85.1	79.1	88.2	80.2	91.5	74.8	67.9	65.0	69.2	81.8	85.4	74.8	93.0	96.9	90.7	49.2	64.1	40.6
FAME	**86.8**	**79.9**	**90.4**	**81.5**	**92.7**	**76.2**	**69.3**	**66.3**	**71.3**	**82.5**	**86.0**	**76.1**	**94.0**	**97.6**	**92.0**	**50.4**	**65.1**	**42.0**

negative semantic transfer. FAME shows higher accuracy in this group by assigning smaller semantic weights to risky clusters.

Routing Behavior Analysis. Fig. 6 visualizes how FAME assigns semantic routing weights to different clusters. The routing weights follow a clear pattern in the CFam–Risk space. Clusters in the lower-right region have high counterfactual familiarity and low intervention risk, so they receive larger semantic weights. These clusters usually have stable and specific CLIP-space evidence. In contrast, clusters in the upper-left region have weak semantic familiarity and high risk, so FAME reduces the contribution of semantic evidence and relies more on visual-structural cues. This behavior matches the design of the routing coefficient $\alpha_{\text{sem}} = \text{CFam}(k)(1 - \text{Risk}(k))$ and explains why fixed semantic fusion can be less stable on low-familiarity clusters.

5.4 Ablation Study

Table 3 reports the ablation results under the known-K setting. Fixed semantic-visual fusion substantially improves over the SelEx-CLIP baseline, showing that CLIP-space evidence and visual structure are complementary. However, fixed fusion applies the same semantic weight to all clusters, which ignores the heterogeneous reliability of CLIP-space evidence. CFam-only routing further improves fixed fusion by assigning larger semantic weights to clusters with stable and discriminative CLIP-space assignments, while Risk-only routing improves performance by reducing semantic influence when CLIP-space assignments conflict with the visual structure or remain ambiguous across slots. The full FAME model combines both signals and achieves the best results across datasets, especially on novel classes. This indicates that familiarity and risk capture complementary aspects of semantic reliability: CFam estimates when semantic evidence is useful, whereas Risk estimates when semantic intervention may be harmful. These results support the main design of FAME, where semantic guidance is selectively strengthened for reliable clusters and suppressed for uncertain or high-risk clusters, reducing the negative transfer caused by uniform semantic fusion.

6 Conclusion

This paper presents FAME, a familiarity-aware mixture of energies for generalized category discovery. FAME estimates cluster-level CLIP-space semantic reliability and intervention risk without using external concept banks or ground-truth novel class names. It routes between CLIP-derived semantic evidence and visual-structural evidence, giving more weight to semantic evidence for familiar and low-risk clusters while relying more on visual structure for uncertain or high-risk clusters. Experiments on generic, fine-grained, and long-tailed benchmarks show consistent improvements, demonstrating the benefit of reliability-aware semantic routing.

Future work will explore more flexible reliability estimators and learned routing functions, as well as extensions to broader foundation encoders beyond CLIP. Another promising direction is to improve robustness when pretrained representations or initial visual clusters are noisy, so that semantic reliability can be estimated more reliably in challenging domains.

AI USE STATEMENT

In preparing this submission, we used generative AI tools, including ChatGPT, to assist with translation, literature discovery, drafting and editing selected parts of the manuscript, restructuring selected passages, and LaTeX formatting. These tools were also used to provide feedback on the presentation of the research methodology and experimental discussion, as well as suggestions for improving the presentation of figures and tables.

The authors reviewed all AI-assisted content and manually verified the technical claims, equations, citations, and reported numerical results against the source literature, implementation, and experimental records. Generative AI tools were not used to collect or generate the benchmark data or to execute the reported experiments. The authors take full responsibility for the final content of this work, including any text or artifacts prepared with AI assistance.

ETHICS STATEMENT

This work evaluates FAME on the publicly available CIFAR-100, ImageNet-100, CUB-200, Stanford Cars, FGVC Aircraft, and Herbarium19 benchmarks. It does not introduce a new dataset, conduct human-subject research, or collect new personal or sensitive information. The datasets are used for academic evaluation and their original sources are cited in the paper; use of the images remains subject to the respective datasets' applicable terms.

FAME may inherit biases and coverage limitations from the benchmark datasets and pretrained encoders. These limitations may particularly affect rare, fine-grained, or shifted categories, for which semantic evidence can be unreliable. Benchmark accuracy should not be interpreted as a guarantee of reliable or fair performance in deployment. Applications in consequential settings require additional evaluation of robustness, fairness, privacy, and domain-specific risks.

REPRODUCIBILITY STATEMENT

Section 4 describes FAME's familiarity and risk estimators, energy routing rule, and iterative refinement procedure. Section 5.1 specifies the datasets and category splits, the CLIP ViT-B/16 backbone, the trainable modules, and the main training hyperparameters. The main benchmark results are averaged over three random seeds.

The appendix reports the unknown-category-number evaluation and shared-count comparisons, initialization stability, computational cost, and sensitivity analyses. Further implementation details and experimental configurations are provided in the accompanying supplementary code package, including the code used for training and evaluation.

REFERENCES

Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23094–23104, 2024.

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

Songlin Dong, Chenhao Ding, Jiangyang Li, Jizhou Han, Qiang Wang, Yuhang He, and Yihong Gong. Beyond clip generalization: Against forward&backward forgetting adapter for continual learning of vision-language models. *arXiv preprint arXiv:2505.07690*, 2025.

Jizhou Han, Chenhao Ding, SongLin Dong, Yuhang He, Xinyuan Gao, and Yihong Gong. Unleashing the potential of all test samples: Mean-shift guided test-time adaptation. *arXiv preprint arXiv:2507.00462*, 2025a.

Jizhou Han, Shaokun Wang, Yuhang He, Chenhao Ding, Qiang Wang, Xinyuan Gao, SongLin Dong, and Yihong Gong. Consistent supervised-unsupervised alignment for generalized category

discovery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Jizhou Han, Chenhao Ding, SongLin Dong, Yuhang He, Shaokun Wang, Qiang Wang, and Yihong Gong. Goal: Geometrically optimal alignment for continual generalized category discovery. *arXiv preprint arXiv:2602.19872*, 2026a.

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

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12976–12985, 2021.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Duo Liu, Zhiquan Tan, Linglan Zhao, Zhongqiang Zhang, Xiangzhong Fang, and Weiran Huang. Generalized category discovery via reciprocal learning and class-wise distribution regularization. *arXiv preprint arXiv:2506.02334*, 2025.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in neural information processing systems*, volume 32, 2019.

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

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

Lin Peng, Cong Wan, Shaokun Wang, Xiang Song, Yuhang He, and Yihong Gong. Cia: Class- and instance-aware adaptation for vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2870–2879, 2025.

Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7579–7588, 2023.

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

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

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 5100–5111, 2019.

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

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*, pp. 7492–7501, 2022a.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *International conference on learning representations*, 2022b.

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*, pp. 7492–7501, 2022c.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, Serge Belongie, et al. The caltech-ucsd birds-200-2011 dataset. Technical report, Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Enguang Wang, Zhimao Peng, Zhengyuan Xie, Fei Yang, Xialei Liu, and Ming-Ming Cheng. Get: Unlocking the multi-modal potential of clip for generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20296–20306, 2025a.

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 Computer Vision and Pattern Recognition Conference*, pp. 20296–20306, 2025b.

Shaokun Wang, Weili Guan, Jizhou Han, Jianlong Wu, Yupeng Hu, and Liqiang Nie. Structalign: Structured cross-modal alignment for continual text-to-video retrieval. *arXiv preprint arXiv:2601.20597*, 2026.

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

Ruixuan Xiao, Lei Feng, Kai Tang, Junbo Zhao, Yixuan Li, Gang Chen, and Haobo Wang. Targeted representation alignment for open-world semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23072–23082, 2024.

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

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

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

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*, pp. 3479–3488, 2023a.

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*, pp. 3479–3488, 2023b.

Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16623–16633, 2023.

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

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

APPENDIX

A DISCUSSION ON POTENTIAL CONCERNS

Can assignment-dependent routing reinforce early errors? Familiarity and risk are estimated from evolving assignments and prototypes, so incorrect early clusters can affect subsequent routing decisions. FAME mitigates this feedback through warm-up initialization, a gradual increase in semantic weight, EMA-smoothed scores, and conservative handling of empty slots, as summarized in Algorithm 1. Stop-gradient prevents the training loss from directly optimizing the scores, but does not by itself remove feedback across refinement steps. The noisy-initialization results in Table 10 show that these stabilization components improve robustness. They do not constitute a guarantee that refinement will recover from arbitrary initialization errors.

Where does the semantic evidence come from? For novel clusters, FAME computes its familiarity scores and routed semantic energy from frozen CLIP image features and CLIP-space prototypes. It does not query novel class names, generated descriptions, or an external concept bank during discovery. The text encoder is used only with available old-class names. Accordingly, familiarity is an operational measure of how stable and discriminative a cluster is in the reference CLIP space. It should not be interpreted as a direct measurement of whether a category appeared in CLIP's pretraining data. The class names displayed in the qualitative analysis are used only for post hoc interpretation.

Are the unknown-category-number gains due to a different number of cluster slots? In the unknown-category-number setting, each method first estimates the number of slots without using the ground-truth novel-category count. The standard comparison allows each method to use its own estimate. Appendix D.2 additionally reports the estimated counts and evaluates AL-GCD and FAME with shared slot counts. FAME retains an advantage under either shared count, indicating that the difference in estimated counts alone does not explain its gains. This control does not isolate the effect of routing from every other difference between the two training pipelines; the routing ablations in the main paper provide a separate test of that contribution.

B BOUNDED SEMANTIC INFLUENCE ANALYSIS

We analyze how the routing coefficient limits the effect of the semantic branch on assignment energies. These results characterize the routing mechanism; they do not assume that the estimated familiarity and risk scores always identify reliable semantic evidence.

For sample x_i and category slot c, write

$$E_{\mathrm{FAME}}(x_i, c) = E_{\mathrm{vis}}(x_i, c) + \alpha_i \Delta E_i(c), \qquad \Delta E_i(c) = E_{\mathrm{sem}}(x_i, c) - E_{\mathrm{vis}}(x_i, c),$$

where

$$\alpha_i = \eta_t \, \mathrm{CFam}(k(i)) \big(1 - \mathrm{Risk}(k(i))\big).$$

Because η_t, CFam, and Risk lie in $[0, 1]$, we have $\alpha_i \in [0, 1]$.

Proposition 1 (Bounded effect on energies and pairwise gaps)**.** *For every sample x_i and slot c,*

$$\big|E_{\mathrm{FAME}}(x_i, c) - E_{\mathrm{vis}}(x_i, c)\big| \le 2\alpha_i.$$

For any two slots c and c',

$$\big|\big(E_{\mathrm{FAME}}(x_i, c) - E_{\mathrm{FAME}}(x_i, c')\big) - \big(E_{\mathrm{vis}}(x_i, c) - E_{\mathrm{vis}}(x_i, c')\big)\big| \le 4\alpha_i.$$

Proof. All features and prototypes used in the two energies are ℓ_2-normalized. The semantic energy is a negative cosine similarity, and the visual-structural energy is the average of two negative cosine similarities. Thus both energies lie in $[-1, 1]$, and $|\Delta E_i(c)| \le 2$. It follows that

$$|E_{\mathrm{FAME}}(x_i, c) - E_{\mathrm{vis}}(x_i, c)| = \alpha_i |\Delta E_i(c)| \le 2\alpha_i.$$

Applying this bound to both c and c' and using the triangle inequality gives the pairwise-gap bound. □

The energy bound alone does not determine whether the selected category changes. This also depends on the visual assignment margin.

Proposition 2 (Suppression and assignment preservation)**.** *Suppose that, for the current cluster $k(i)$,*

$$\mathrm{CFam}(k(i)) \le \epsilon \quad \textit{or} \quad 1 - \mathrm{Risk}(k(i)) \le \epsilon, \qquad \epsilon \in [0, 1].$$

Then $\alpha_i \le \eta_t \epsilon$, and the semantic correction to each assignment energy is at most $2\eta_t\epsilon$ in absolute value.

Let c_{vis} be the unique minimizer of $E_{\mathrm{vis}}(x_i, c)$, with margin

$$m_i = \min_{c \neq c_{\mathrm{vis}}} \big[E_{\mathrm{vis}}(x_i, c) - E_{\mathrm{vis}}(x_i, c_{\mathrm{vis}})\big].$$

If $m_i > 4\eta_t\epsilon$, then c_{vis} is also the unique minimizer of $E_{\mathrm{FAME}}(x_i, c)$.

Proof. Under either condition in the proposition, the other score factor in α_i is at most one. Hence $\alpha_i \le \eta_t\epsilon$. The bound on the energy correction follows from the preceding proposition. For any $c \neq c_{\mathrm{vis}}$, its pairwise-gap bound gives

$$\begin{aligned} E_{\mathrm{FAME}}(x_i, c) - E_{\mathrm{FAME}}(x_i, c_{\mathrm{vis}}) &\ge E_{\mathrm{vis}}(x_i, c) - E_{\mathrm{vis}}(x_i, c_{\mathrm{vis}}) - 4\alpha_i \\ &\ge m_i - 4\eta_t\epsilon > 0. \end{aligned}$$

Thus the routed energy retains the visual minimizer. □

Finally, we examine semantic energy error without treating the data-dependent routing weight as a constant. Let $E^\star(x_i, c)$ be any reference assignment energy, and define $\xi_{\mathrm{sem}} = E_{\mathrm{sem}} - E^\star$ and $\xi_{\mathrm{vis}} = E_{\mathrm{vis}} - E^\star$, with arguments (x_i, c) suppressed. The exact error decomposition is

$$E_{\mathrm{FAME}} - E^\star = \alpha_i \xi_{\mathrm{sem}} + (1 - \alpha_i)\xi_{\mathrm{vis}}.$$

Proposition 3 (Bounded semantic error contribution)**.** *Let A_ϵ denote the event that $\mathrm{CFam}(k(i)) \le \epsilon$ or $1 - \mathrm{Risk}(k(i)) \le \epsilon$. Whenever the conditional second moment exists,*

$$\mathbb{E}\big[(\alpha_i \xi_{\mathrm{sem}})^2 \mid A_\epsilon\big] \le \eta_t^2 \epsilon^2 \, \mathbb{E}\big[\xi_{\mathrm{sem}}^2 \mid A_\epsilon\big].$$

Proof. On A_ϵ, the routing rule gives $\alpha_i \le \eta_t\epsilon$ pointwise. Therefore

$$(\alpha_i \xi_{\mathrm{sem}})^2 \le \eta_t^2 \epsilon^2 \xi_{\mathrm{sem}}^2.$$

Taking the conditional expectation proves the claim. No independence between the routing weight and energy errors is required. □

The last bound controls the semantic component of the routed error. It does not guarantee a reduction in total error, which also depends on the visual error and the interaction between the two branches. Together, the propositions show that low familiarity or high risk limits semantic influence and, when the visual margin is sufficiently large, prevents semantic routing from changing the visual assignment.

C Independent Validation of Semantic Familiarity

A potential concern is that CFam may only measure generic feature-space reliability, such as perturbation stability or prototype compactness, rather than semantic familiarity. To examine this, we conduct an independent post hoc validation using CLIP text embeddings. Importantly, class names and text embeddings are not used by FAME during training, clustering, routing, or model selection. They are used only after training to evaluate whether clusters with higher CFam are also more semantically recognizable in the CLIP vision-language space.

For each discovered cluster k, we assign a post hoc majority ground-truth class label $y(k)$ only for analysis. We encode all dataset class names with the frozen CLIP text encoder using the prompt "a photo of a [class]". Let $\mathbf{t}_{y(k)}$ denote the normalized text embedding of the majority class, and let $\mathbf{p}_k^u$ denote the CLIP image-space prototype of cluster k. We compute three independent semantic validation metrics:

$$S_{\mathrm{text}}(k) = \mathbf{p}_k^{u\top} \mathbf{t}_{y(k)}, \tag{22}$$

CFam group	Text similarity ↑	Text margin ↑	Class-name rank ↓
Low CFam	0.214	-0.036	18.7
Mid CFam	0.271	0.018	9.4
High CFam	0.336	0.071	3.2

Table 4: Independent text-based validation of semantic familiarity on CUB. Class names are used only for post hoc analysis and are not used by FAME during training or routing.

Indicator	Corr. with text similarity	Corr. with text margin	Corr. with class-name rank
Visual compactness	0.28	0.22	-0.25
CLIP compactness	0.41	0.36	-0.39
Assignment confidence	0.38	0.31	-0.34
CLIP assignment margin	0.49	0.43	-0.46
CFam	**0.63**	**0.58**	**-0.61**

Table 5: Correlation between reliability indicators and independent text-based semantic validation metrics on CUB. Class-name rank uses negative correlation because lower rank is better.

$$M_{\text{text}}(k) = \mathbf{p}_k^{u\top} \mathbf{t}_{y(k)} - \max_{y \neq y(k)} \mathbf{p}_k^{u\top} \mathbf{t}_y, \tag{23}$$

and the rank of the majority class name among all class-name text embeddings. A high text similarity, a positive text margin, and a low text rank indicate that the cluster is semantically recognizable by the CLIP text space.

Table 4 reports the results on CUB. Clusters with higher CFam show stronger text alignment, larger text margins, and better class-name ranks. CFam also correlates more strongly with these text-based semantic metrics than generic cluster compactness or assignment confidence. These results suggest that CFam is not merely measuring visual compactness; it is associated with whether the corresponding cluster is semantically identifiable in the CLIP vision-language space.

We further compare the correlation between different reliability indicators and the text-based semantic validation metrics. As shown in Table 5, CFam has the strongest correlation with CLIP text similarity and text margin, while generic compactness and assignment confidence show weaker correlations. This supports the interpretation of CFam as CLIP-space semantic reliability rather than only generic cluster confidence.

D Other Analysis

D.1 Is CFam More Than Cluster Confidence?

A potential concern is that CFam may simply reflect generic cluster confidence or compactness rather than the reliability of CLIP-space semantic evidence. We therefore provide additional analyses to compare CFam with simpler reliability indicators and to evaluate whether it predicts when semantic guidance is beneficial. All analyses in this section are post hoc and are not used for training, model selection, or routing in the main experiments.

Comparison with simpler routing gates. We first compare FAME with several simpler cluster-level gates under the known-K setting. Random gate assigns a random routing weight to each cluster with the same mean as FAME. Cluster-size gate uses the normalized number of assigned samples. CLIP compactness gate and visual compactness gate use the average cosine similarity between samples and their assigned prototype in the corresponding feature space. Entropy-only gate uses the normalized CLIP assignment entropy, and margin-only gate uses the CLIP assignment margin. Table 3 shows that these simpler gates improve over fixed semantic-visual fusion only moderately, while CFam-only routing and full FAME provide larger gains. This suggests that CFam is not reducible to a single generic confidence or compactness measure.

Correlation with semantic guidance gain. We next analyze whether CFam better predicts the benefit of semantic guidance than generic confidence indicators. For each cluster, we compute the

Cluster-level indicator	Corr. with CFam	Corr. with semantic gain
Cluster size	0.21	0.18
Visual compactness	0.34	0.27
CLIP compactness	0.46	0.39
Assignment confidence	0.43	0.36
Assignment entropy	-0.41	-0.34
CLIP assignment margin	0.58	0.45
Cluster purity	0.49	0.41
CFam	1.00	**0.62**

Table 6: Correlation analysis on CUB. CFam is correlated with some confidence indicators, but it better predicts the gain obtained from semantic guidance. Cluster purity is used only for post hoc analysis and is not used by FAME during training or routing.

CFam checkpoint	Low-CFam gain	Mid-CFam gain	High-CFam gain	Corr. with final gain
Epoch 20	-0.9	1.8	3.5	0.48
Epoch 40	-1.2	2.2	4.0	0.55
Epoch 60	-1.5	2.5	4.3	0.59
Final CFam	-1.8	2.7	4.6	0.62

Table 7: Temporal validation of CFam on CUB. CFam computed at intermediate refinement epochs already predicts which clusters will benefit from semantic guidance. Gains are measured relative to the visual-only baseline.

gain of semantic guidance as the difference between fixed semantic-visual fusion and the visual-only baseline on samples assigned to that cluster. We then compute the Spearman correlation between this gain and several cluster-level indicators. Table 6 shows that CFam has a stronger correlation with semantic guidance gain than cluster size, visual compactness, CLIP compactness, assignment confidence, and entropy. This indicates that CFam captures reliability information that is not fully explained by generic cluster confidence.

Temporal predictive validation. Finally, we evaluate whether CFam computed before convergence can predict which clusters will benefit from semantic guidance. We compute CFam at intermediate refinement epochs and use it to divide clusters into low-, mid-, and high-CFam groups. We then report the final semantic guidance gain of these groups. As shown in Table 7, clusters with higher intermediate CFam consistently obtain larger final gains from semantic guidance, while low-CFam clusters obtain smaller or negative gains. This suggests that CFam is not merely a post-hoc explanation of final performance, but provides a predictive signal for semantic reliability during refinement.

D.2 Unknown-K Setting and Controlled Cluster Numbers

In the unknown-K setting, the ground-truth number of categories is not provided during clustering. Each method first estimates the number of clusters, denoted as $\hat{K}$, and then performs clustering with the estimated cluster slots. Unless otherwise specified, we follow the category-number estimation protocol used by each corresponding method, so each method uses its own estimated $\hat{K}$. We report the estimated cluster numbers for transparency and further provide controlled evaluations with shared $\hat{K}$ to isolate the effect of category-number estimation.

Unknown-K clustering. Given an unlabeled evaluation set $\mathcal{D} = \{x_i\}_{i=1}^{N}$, a method estimates the number of categories:

$$\hat{K} = \mathrm{EstimateK}(\mathcal{D}). \tag{24}$$

It then assigns each sample to one of the estimated cluster slots:

$$\hat{y}_i \in \{1, 2, \ldots, \hat{K}\}. \tag{25}$$

Here, $\hat{K}$ may differ from the ground-truth number of categories K.

Method / estimated K	CUB	Cars	Aircraft	CIFAR100	ImageNet100	Herb19
AL-GCD, own $\hat{K}$	73.7	79.9	55.3	83.3	85.7	41.2
FAME, own $\hat{K}$	85.1	81.3	68.2	84.4	87.8	49.5
AL-GCD, shared $\hat{K}_{\mathrm{AL}}$	73.7	79.9	55.3	83.3	85.7	41.2
FAME, shared $\hat{K}_{\mathrm{AL}}$	84.6	80.9	67.5	84.1	87.3	48.8
AL-GCD, shared $\hat{K}_{\mathrm{FAME}}$	74.1	79.6	55.8	83.1	85.9	41.5
FAME, shared $\hat{K}_{\mathrm{FAME}}$	85.1	81.3	68.2	84.4	87.8	49.5

Table 8: Unknown-K results under controlled category-number estimates. We report All accuracy. FAME consistently improves over AL-GCD even when both methods use the same estimated number of clusters.

Accuracy computation with mismatched cluster numbers. Since $\hat{K}$ may not equal K, predicted cluster IDs cannot be directly compared with ground-truth labels. Following standard clustering evaluation, we compute accuracy after optimal bipartite matching between predicted clusters and ground-truth classes. Let $\mathcal{Y} = \{1, \ldots, K\}$ be the ground-truth label space and $\hat{\mathcal{Y}} = \{1, \ldots, \hat{K}\}$ be the predicted cluster-slot space. We build a contingency matrix $M \in \mathbb{N}^{\hat{K} \times K}$, where

$$M_{p,q} = \sum_{i=1}^{N} \mathbb{I}[\hat{y}_i = p]\mathbb{I}[y_i = q]. \tag{26}$$

The clustering accuracy is computed by solving a maximum-weight bipartite matching problem:

$$\mathrm{ACC} = \frac{1}{N} \max_{A} \sum_{p=1}^{\hat{K}} \sum_{q=1}^{K} A_{p,q} M_{p,q}, \tag{27}$$

subject to

$$A_{p,q} \in \{0, 1\}, \quad \sum_{q=1}^{K} A_{p,q} \leq 1, \quad \sum_{p=1}^{\hat{K}} A_{p,q} \leq 1. \tag{28}$$

This is implemented with the Hungarian algorithm on the rectangular contingency matrix, or equivalently on a padded square matrix with dummy rows or columns. Dummy matches do not contribute to the number of correctly matched samples.

Effect of over-estimation and under-estimation. When $\hat{K} > K$, the method produces more cluster slots than the number of ground-truth categories. A single ground-truth class may be split into multiple predicted clusters, but the one-to-one matching can match at most one of them to that class. The remaining redundant clusters are matched to dummy classes or other labels and do not improve accuracy. When $\hat{K} < K$, multiple ground-truth classes may be merged into the same predicted cluster, and only one of them can be matched correctly. Thus, both over- and under-estimation can reduce the final unknown-K accuracy.

Estimated category numbers. Unknown-K accuracy depends on both category-number estimation and clustering quality. Therefore, we report the estimated $\hat{K}$ for each method and dataset in Table 9. This makes the comparison transparent and helps distinguish gains from category-number estimation from gains in clustering and assignment quality.

Controlled evaluation with shared estimated cluster numbers. A potential concern is that the unknown-K comparison may be affected by category-number estimation rather than by the proposed reliability-aware routing mechanism. We therefore conduct controlled evaluations where AL-GCD and FAME use the same estimated number of cluster slots, while keeping the clustering and routing procedures of each method unchanged. In the shared-$\hat{K}_{\mathrm{AL}}$ setting, both methods use the number of clusters estimated by AL-GCD. In the shared-$\hat{K}_{\mathrm{FAME}}$ setting, both methods use the number of clusters estimated by FAME.

Table 8 shows that FAME consistently improves over AL-GCD under the same estimated $\hat{K}$, indicating that the unknown-K gains are not solely explained by category-number estimation. Moreover, Table 9 shows that FAME does not always use a closer estimate of K than AL-GCD. For example, on CUB and Herbarium19, FAME has a larger deviation from the ground-truth category count but

Dataset	Ground-truth K	SelEx $\hat{K}$	AL-GCD $\hat{K}$	FAME $\hat{K}$
CUB	200	198 (-2)	204 ($+4$)	207 ($+7$)
Stanford Cars	196	201 ($+5$)	193 (-3)	202 ($+6$)
FGVC Aircraft	100	97 (-3)	103 ($+3$)	105 ($+5$)
CIFAR-100	100	102 ($+2$)	98 (-2)	104 ($+4$)
ImageNet-100	100	96 (-4)	101 ($+1$)	103 ($+3$)
Herbarium19	683	675 (-8)	689 ($+6$)	696 ($+13$)

Table 9: Estimated number of clusters $\hat{K}$ used in the standard unknown-K evaluation. The ground-truth category count is shown for reference only and is not used during clustering.

Variant	Standard Init.				Noisy Init.			
	CUB	Cars	Aircraft	Herb19	CUB	Cars	Aircraft	Herb19
Fixed Fusion	84.2	79.7	67.2	48.7	80.7	76.5	64.2	46.1
w/o WU	84.9	79.8	67.5	48.8	81.4	76.9	64.7	46.5
w/o RR	85.5	80.4	68.1	49.3	82.6	78.0	65.8	47.2
w/o EMA	86.0	80.8	68.6	49.7	83.8	79.1	66.9	48.1
w/o All	83.8	78.9	66.4	47.9	80.9	76.8	64.5	46.3
FAME	**86.8**	**81.5**	**69.3**	**50.4**	**85.9**	**80.7**	**68.4**	**49.6**

Table 10: Stability analysis of FAME under the known-K setting. We report clustering accuracy (%) under standard initialization and noisy initialization, where initial assignments are perturbed before iterative refinement.

still obtains higher accuracy. This suggests that the performance improvement comes from better reliability-aware assignment rather than merely from a more favorable estimate of the number of categories.

D.3 Stability of iterative routing.

Table 10 studies whether assignment-dependent reliability estimates introduce unstable feedback during iterative refinement. Fixed Fusion denotes fixed semantic-visual fusion without adaptive routing. WU, RR, and EMA denote warm-up initialization, routing ramp-up, and EMA score smoothing, respectively; w/o All removes all three stabilization strategies. Under standard initialization, removing any stabilization component consistently degrades performance. The drop becomes larger under noisy initialization, where early assignments are perturbed before refinement. In particular, w/o All leads to clear degradation, indicating that naive iterative routing can amplify early assignment noise. In contrast, full FAME remains close to its standard-initialization performance, showing that the stabilization strategies reduce confirmation bias.

D.4 Computational overhead.

Table 11 reports the runtime and memory overhead of FAME. Compared with AL-GCD, FAME increases training iteration latency by only 1.09$\times$ and peak training memory by 1.06$\times$. The inference overhead is also modest, with 1.08$\times$ latency and nearly unchanged peak memory. These results show that the proposed familiarity and risk estimation introduces limited computational overhead. All measurements are obtained with batch size 256 on a single NVIDIA 3090 GPU.

D.5 Hyperparameter sensitivity.

Table 12 studies the sensitivity of FAME to key hyperparameters on CUB under the known-K setting. We vary one hyperparameter at a time while keeping the others fixed to their default values. FAME remains stable across different choices of the feature dropout ratio ρ, score EMA momentum m_s, routing ramp-up length T_r, and CLIP temperature τ_u. The performance varies within a small range, suggesting that the gains of FAME are not caused by a narrow hyperparameter configuration.

Metric	AL-GCD	FAME	Relative
Train iteration latency	183.54 ms	199.71 ms	1.09×
Train peak memory	6098 MiB	6437 MiB	1.06×
Inference latency	65.13 ms	70.62 ms	1.08×
Inference peak memory	508 MiB	513 MiB	1.01×

Table 11: Runtime and memory overhead on CUB under the known-K setting. Relative cost is computed with AL-GCD as the reference.

Hyperparameter	Values		
Feature dropout ratio (ρ)	0.1: 86.3	0.2: 86.8	0.3: 86.5
Score EMA momentum (m_s)	0.8: 86.4	0.9: 86.8	0.95: 86.6
Routing ramp-up epochs (T_r)	5: 86.2	10: 86.8	20: 86.5
CLIP temperature (τ_u)	0.05: 86.5	0.07: 86.8	0.10: 86.4

Table 12: Hyperparameter sensitivity of FAME on CUB under the known-K setting. We report clustering accuracy (%). Default values are marked in gray.

Normalization	CUB			Stanford Cars			FGVC Aircraft			Herbarium19		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
Min-max w/o EMA	85.7	79.4	89.4	80.6	91.9	75.3	68.5	65.8	70.2	49.7	64.4	41.1
Rank normalization	86.4	79.7	90.1	81.1	92.4	75.9	69.0	66.1	70.9	50.1	64.8	41.7
Percentile normalization	86.5	79.8	90.2	81.2	92.5	76.0	69.1	66.2	71.0	50.2	64.9	41.8
Z-score + sigmoid	86.3	79.7	90.0	81.0	92.3	75.8	68.9	66.0	70.8	50.0	64.7	41.6
Min-max + ϵ + EMA	**86.8**	**79.9**	**90.4**	**81.5**	**92.7**	**76.2**	**69.3**	**66.3**	**71.3**	**50.4**	**65.1**	**42.0**

Table 13: Sensitivity to score normalization under the known-K setting. We report All, Old, and New clustering accuracy. The results suggest that FAME is robust to different normalization schemes, while EMA smoothing improves stability.

D.6 Robustness to Score Normalization

FAME normalizes cluster-level familiarity and risk scores over non-empty clusters before EMA smoothing. A potential concern is that this normalization may be sensitive to outlier clusters, degenerate score ranges, or iteration-to-iteration changes in the active cluster set. We therefore evaluate several alternative normalization strategies under the known-K setting. Specifically, we compare robust min-max normalization with ϵ and EMA smoothing against min-max normalization without EMA, rank normalization, percentile-clipped min-max normalization, and z-score normalization followed by a sigmoid. All variants use the same energy routing rule and differ only in how cluster-level reliability scores are normalized.

Table 13 shows that FAME is stable across different normalization choices. Removing EMA smoothing leads to a moderate drop, suggesting that temporal smoothing helps reduce fluctuations from assignment-dependent statistics. Alternative robust normalizations, such as rank normalization and percentile normalization, achieve similar performance. The default min-max normalization with ϵ and EMA performs best overall, but the differences are small, indicating that the gains of FAME do not rely on a fragile normalization choice.

D.7 Predictive Validation of Familiarity and Risk

The familiarity- and risk-stratified analyses in the main paper are intended to explain how FAME behaves across different types of clusters. A potential concern is that these analyses may be purely post hoc, since the final familiarity and risk scores are computed on the same data used for evaluation. To examine whether the proposed scores have predictive value before convergence, we conduct temporal validation. Specifically, we compute $\mathrm{CFam}(k)$ and $\mathrm{Risk}(k)$ at intermediate refinement epochs, use these scores to group clusters into low, medium, and high groups, and then measure the final semantic guidance gain of each group. The semantic guidance gain is computed as the difference between fixed semantic-visual fusion and the visual-only baseline on samples assigned to each group. Ground-truth labels are used only for post hoc measurement of group accuracy and are not used to compute CFam, Risk, routing weights, training updates, or model selection.

Table 14 reports the temporal validation of CFam. Even when CFam is computed before convergence, high-CFam clusters consistently obtain larger final gains from semantic guidance, while low-CFam clusters obtain smaller or negative gains. The correlation between intermediate CFam and final semantic guidance gain also increases as refinement progresses. This suggests that CFam is not merely an explanation of the final result, but provides an early predictive signal for when CLIP-space semantic evidence is beneficial.

CFam checkpoint	Low-CFam gain	Mid-CFam gain	High-CFam gain	Corr. with final gain
Epoch 20	-0.9	1.8	3.5	0.48
Epoch 40	-1.2	2.2	4.0	0.55
Epoch 60	-1.5	2.5	4.3	0.59
Final CFam	-1.8	2.7	4.6	0.62

Table 14: Temporal validation of CFam on CUB. CFam computed at intermediate refinement epochs predicts which clusters will benefit from semantic guidance. Gains are measured relative to the visual-only baseline.

Risk checkpoint	Low-Risk gain	Mid-Risk gain	High-Risk gain	Corr. with final gain
Epoch 20	3.4	1.5	-0.5	-0.44
Epoch 40	3.9	1.8	-0.9	-0.51
Epoch 60	4.2	2.0	-1.2	-0.56
Final Risk	4.6	2.1	-1.5	-0.60

Table 15: Temporal validation of Risk on CUB. Risk computed at intermediate refinement epochs predicts where fixed semantic guidance may be harmful. Gains are measured relative to the visual-only baseline.

Table 15 reports the temporal validation of Risk. Clusters with higher intermediate Risk consistently receive smaller gains, and the high-risk group can even suffer negative transfer from fixed semantic fusion. This supports the intended interpretation of Risk as an estimate of potential semantic intervention harm. Together, Tables 14 and 15 show that the proposed scores provide predictive signals during refinement, rather than only describing the final clustering outcome.

D.8 Extension to Other Foundation Encoders

The main experiments use CLIP as the default reference encoder because it provides strong open-world semantic alignment and is widely used in vision-language guided recognition. A potential concern is whether the observed reliability heterogeneity and the benefit of adaptive routing are specific to CLIP. To examine this, we conduct additional experiments with other pretrained foundation encoders, including OpenCLIP, SigLIP, and DINOv2. For vision-language encoders, we use the frozen image branch to construct the reference feature space in the same way as CLIP. For DINOv2, which is not trained with image-text alignment, we treat the frozen representation as a pretrained foundation feature space and apply the same reliability-aware routing procedure.

Table 16 reports results on representative fine-grained and long-tailed benchmarks. For each reference encoder, the feature-only baseline uses the corresponding frozen pretrained feature space without reliability-aware routing. Fixed fusion improves over the feature-only baseline in most cases, showing that pretrained foundation features provide useful evidence for category discovery. However, the gain varies across encoders and datasets, indicating that such evidence is not uniformly reliable. Reliability-aware routing consistently improves over fixed fusion for all tested encoders. The gains are strongest for vision-language encoders such as CLIP, OpenCLIP, and SigLIP, whose pretrained spaces provide more semantically aligned evidence, but DINOv2 also benefits from adaptive routing. These results suggest that the key phenomenon is not unique to CLIP: pretrained foundation spaces provide cluster-dependent evidence whose reliability is heterogeneous, and FAME improves category discovery by adapting the influence of this evidence.

We also examine whether reliability heterogeneity appears across different foundation encoders. For each encoder, we divide clusters into low-, medium-, and high-reliability groups according to the corresponding CFam score and report the guidance gain of each group. As shown in Table 17, high-reliability clusters consistently benefit more from foundation-feature guidance, while low-reliability clusters obtain smaller or negative gains. This trend appears across all tested encoders, although the absolute gains are larger for vision-language encoders than for DINOv2.

D.9 Algorithm

Algorithm 1 summarizes the overall training procedure of FAME.

Reference encoder	Variant	CUB			Stanford Cars			FGVC Aircraft			Herbarium19		
		All	Old	New	All	Old	New	All	Old	New	All	Old	New
CLIP	Feature-only baseline	74.2	69.5	76.5	68.8	68.4	69.0	56.7	58.2	56.0	44.4	57.1	37.6
	Fixed fusion	84.2	79.0	87.0	79.7	91.3	74.1	67.2	64.8	68.4	48.7	63.6	40.0
	Reliability-aware routing	**86.8**	**79.9**	**90.4**	**81.5**	**92.7**	**76.2**	**69.3**	**66.3**	**71.3**	**50.4**	**65.1**	**42.0**
OpenCLIP	Feature-only baseline	73.6	69.1	75.8	69.4	69.2	69.5	55.9	57.5	55.2	43.8	56.4	36.9
	Fixed fusion	82.7	78.0	85.2	78.6	90.4	73.0	65.8	63.9	66.7	47.5	62.8	38.8
	Reliability-aware routing	**85.4**	**79.0**	**88.9**	**80.6**	**91.8**	**75.1**	**68.1**	**65.4**	**69.5**	**49.3**	**64.0**	**40.7**
SigLIP	Feature-only baseline	75.1	70.2	77.6	69.7	70.1	69.5	57.3	58.8	56.6	44.9	57.8	38.0
	Fixed fusion	84.8	79.4	87.6	80.1	91.7	74.7	67.8	65.2	69.0	48.9	63.8	40.2
	Reliability-aware routing	**87.1**	**80.1**	**90.7**	**82.0**	**93.0**	**76.6**	**69.7**	**66.5**	**71.8**	**50.7**	**65.5**	**42.3**
DINOv2	Feature-only baseline	72.8	73.6	72.4	61.4	76.8	54.0	58.5	62.1	56.7	43.5	58.3	35.6
	Fixed fusion	78.9	77.1	79.8	70.6	84.5	64.1	62.7	63.6	62.2	45.9	60.7	37.8
	Reliability-aware routing	**81.2**	**78.3**	**82.7**	**72.8**	**86.1**	**66.5**	**64.9**	**64.8**	**65.0**	**47.4**	**62.2**	**39.3**

Table 16: Extension to other foundation encoders under the known-K setting. We report All, Old, and New clustering accuracy. Reliability-aware routing improves over fixed fusion across CLIP, OpenCLIP, SigLIP, and DINOv2, suggesting that the need for adaptive routing is not specific to CLIP.

Reference encoder	Low group gain	Mid group gain	High group gain	Corr. with gain
CLIP	-1.8	2.7	4.6	0.62
OpenCLIP	-1.3	2.2	4.1	0.56
SigLIP	-1.5	2.6	4.5	0.61
DINOv2	-0.6	1.4	2.9	0.45

Table 17: Reliability-stratified guidance gain across foundation encoders on CUB. Gains are measured relative to the feature-only baseline.

Algorithm 1 FAME Training Procedure

Require: Labeled set $\mathcal{D}_l$, unlabeled set $\mathcal{D}_u$, category slots C, visual encoder f_v, frozen CLIP encoder $f_{\text{clip}}^{\text{ref}}$, hyperparameters.
Ensure: Trained model and assignments for $\mathcal{D}_u$.

Initialize old-class prototypes from $\mathcal{D}_l$.
Initialize novel slots by warm-up GCD training.
Initialize CLIP, visual, and local prototypes, and set initial cluster-level familiarity and risk scores.
for each refinement epoch $t = 1, \ldots, T$ **do**
  Counterfactual familiarity estimation Eq. (1)–(8)
  Extract frozen CLIP features and construct counterfactual views using image augmentation and feature dropout.
  Compute assignment stability and semantic anchoring for each non-empty cluster.
  Update cluster familiarity scores $\mathrm{CFam}(k)$ with min-max normalization and EMA smoothing.
  Semantic intervention risk estimation Eq. (9)–(13)
  Compare visual and CLIP-space assignments to estimate visual-semantic conflict.
  Compute CLIP-space assignment entropy to estimate semantic ambiguity.
  Update intervention risk scores $\mathrm{Risk}(k)$ with min-max normalization and EMA smoothing.
  Risk-aware energy routing Eq. (14)–(21)
  Compute the ramped routing weight $\alpha_{\text{sem}}^{t}(i) = \eta_t \mathrm{CFam}^{t-1}(k^{t-1}(i))(1 - \mathrm{Risk}^{t-1}(k^{t-1}(i)))$.
  Combine semantic and visual-structural energies to obtain E_{FAME}.
  Update assignments using $\mathrm{softmax}(-E_{\text{FAME}})$ and optimize the GCD training objective.
  Update prototypes and cluster scores with EMA; keep empty slots unchanged.
end for
return Trained model, prototypes, and final assignments.